DOCUMENT RESUME

ED 317 590 TM 014 655

AUTHOR Lewy, Arieh
TITLE Issues in Curriculum Evaluation.
INSTITUTION Ministry of Education and Culture, Jerusalem (Israel).
REPORT NO ISBN-965-262-015-5
PUB DATE 88
NOTE 134p.
PUB TYPE Books (010) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Curriculum Development; *Curriculum Evaluation; *Educational Assessment; Evaluation Methods; Evaluation Problems; *Evaluation Utilization; *Evaluators; Foreign Countries; *Formative Evaluation; *Program Evaluation; Qualitative Research; Summative Evaluation

ABSTRACT

The chapters in this book address four major concerns in the field of curriculum evaluation. The first chapter focuses on the formative evaluation of educational programs, concentrating on the link between program development and formative evaluation. The second chapter--"Evaluating Curriculum Implementation"--presents a classification scheme for evaluation studies, and discusses the evaluation of implementation of educational programs. The third chapter considers quantitative and qualitative evaluation and recommends the use of both types of methods. The fourth chapter--"Utilizing Evaluation Results"--reviews factors affecting utilization and problems associated with the use of evaluation results. The sequence of these chapters reflects shifts in the locus of interest among evaluators. The distinction between summative and formative evaluation coincided with the emergence of curriculum evaluation as a separate field in the 1950s. Interest in implementation evaluation increased in the 1960s, when widespread use of qualitative models in curriculum evaluation also began. Evaluation utilization became an increasing concern in the 1970s and remains the subject of much interest. Eleven figures are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317590

ISSUES IN CURRICULUM EVALUATION

נושאים בהערכה של תוכניות לימודים

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

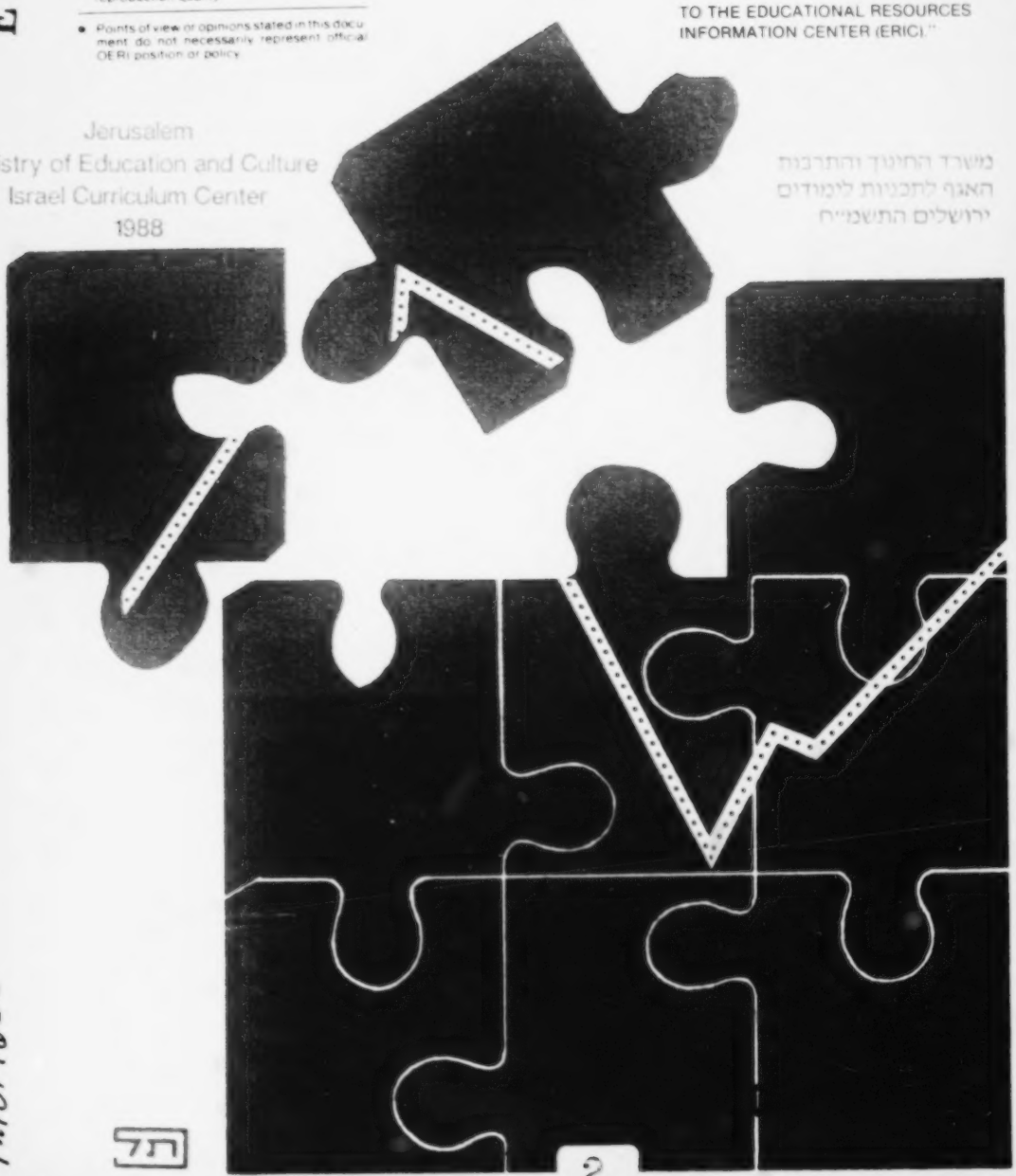
ARIEH LEWY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Arieh Lewy

Jerusalem
Ministry of Education and Culture
Israel Curriculum Center
1988

משרד החינוך והתרבות
האגף לתכניות לימודים
ירושלים התשמ"ח



TM014655

ISSUES IN CURRICULUM EVALUATION

Arieh Lewy

Jerusalem
Ministry of Education and Culture
Israel Curriculum Center
1988

Issues in Curriculum Evaluation

ISBN 965 – 262 – 015 – 5

Table of Contents

Introduction		1
Chapter 1.	**Formative Evaluation of Educational Programs: The State of Art**	3
	The Link between Between Program Development and Formative Evaluation	5
	Phases of Program Development	8
	Pre-development phase	10
	Production of prototype components and their assembly	17
	Release of the in-house assembled version	22
	The revision of the curriculum	37
	Concluding remarks	38
Chapter 2.	**Evaluating Curriculum Implementation**	45
	A Classification Scheme for Evaluation Studies Dealing with Curriculum Implementation	48
	Deciptive Dimensions	49
	The Summative – Formative Dimension of Implementation Evaluation	67
	Measuring Degree of Program Implementation	69
	Implementation as an Independent or as a Dependent Variable	72
	The Openness of the Evaluation System	74
Chapter 3.	**Quantitative and Qualitative Paradigms**	81
	The Quantitative-Qualitative Continuum	84
	Examples of Quantitative and Qualitative Studies	86
	From Disdain to Detente	97
	Why is the Mixing of Methods Recommended?	99
Chapter 4.	**Utilizing Evaluation Results**	105
	Factors Affecting Utilization	109
	Utility Standards	114
	The Complexity of Evaluation Utilization	118
	The Importance of Utilization : An Adversary View	120
	Concluding Remarks	122
Subject index		127
Authors' index		131

List of Tables

1.1	Phases of development and evaluation activities	9
1.2	A ten item checklist	23
1.3	Excerpts from Eash's model	24
1.4	Excerpts from an eclectic checklist	26
1.5	Percentage of correct responses in a sample asample of six schools	32
1.6	Pass and fail for hierarchy pairs of items	36
2.1	Percentage of teachers observed implementing	58
2.2	Levels of Use of Innovation	59
2.3	Potential difficulties	60
2.4	Average time engaged in a specific interaction	64
2.5	Techniques for measuring implementation	69
2.6	Records often produced by educational programs	71
3.1	Clusters of Variables	89
3.2	List of evaluation focuses	93
4.1	Properties affecting evaluation utilization	109
4.2	Utility standards	114
4.3	Report timeliness	115

List of Figures

1.1	Starting a formative evaluation study	7
1.2	Time and freedom of action	8
1.3	An example of a picture and its revision	19
1.4	A hierachical structure of a learning unit	34
1.5	A model of curriculum revision	37
2.1	The field of evaluating curriculum implementation	47
2.2	Mapping sentence of curriculum implementation evaluation	49
2.3	A structural model of variables of implementation	67
2.4	Innovation profile of 30 science teachers according to levels of use on five curricular dimensions	69
3.1	Mapping sentence of curriculum evaluation	81
3.2	Dimensions of variation in multisite multimethod studies	85

Introduction

The articles contained in this book touch on four major concerns in the field of curriculum evaluation. Their sequence in the book reflects shifts in the locus of interest among evaluators.

The distinction made between formative and summative evaluation historically coincides with the emergence of curriculum evaluation as a separate field of disciplined enquiry. In the 1950s numerous theoretical and empirical studies were published which elaborated on, or illustrated the differences between these two archetypes of curriculum evaluation studies. While no change has occurred in the importance attributed to these two study types, it seems that the interest in the differences between such studies has decreased.

Interest in implementation evaluation increased in the 1960s when the products of the new curriculum movement reached the schools. Failure to implement many of the innovative programs created interest in dealing with this topic.

The utilization of qualitative models in curriculum evaluation became widespread in the late 1960s and since then it has been possible to observe a continuous growth in studies of this type.

The first qualitative studies in curriculum evaluation constituted an attempt to substitute the quantitative models, but recently they have been conceived as complementary rather than rival models.

Finally, problems related to utilizing evaluation research became a concern of professionals during the 1970s. Since then, attempts have been made to systematize studies dealing with problems of utilization, to establish codes related to utilization and to increase utilization in practice.

Chapter One

Formative Evaluation of Educational Programs: The State of the Art

The distinction between Formative and Summative Evaluation by Scriven (1967) is one of the most influential innovations in the field of educational evaluation in the last three decades. These forms of evaluation have had tremendous impact on the conceptual and pragmatic aspects of evaluation studies. The use of these terms has become so widespread that it is surprising to discover that they came into use only thirty years ago. Originally the terms were used in the context of curriculum evaluation only, referring to evaluation activities carried out during the process of program development. To justify evaluations of this type Cronbach (1963) claimed:

> "To be influential in course improvement evidence must become available midway in the curriculum development, not in the home stretch, when the developer is naturally reluctant to tear open a supposedly finished body of materials and techniques."
>
> (Cronbach 1963 p. 675)

Cronbach described the characteristic features of evaluation for course improvement (which later became known as Formative Evaluation) and claimed that such evaluation should ascertain what changes the use of certain instructional materials produce in pupils.

> "Outcomes of instruction are multidimensional, and satisfactory investigation will map out the effects of the course along these dimensions separately. To agglomerate many types of post-course performance into a single score is a mistake, because failure to achieve one objective is masked by success in another direction."
>
> (Cronbach 1963 p. 675)

As indicated by Cronbach the flaws of a particular educational program can be eliminated only if the developers obtain separate information about each significant aspect.

This concept of course improvement based on dealing separately with each particular component of program-effect is in glaring contrast with Scriven's (1967) conceptualization of evaluation. Scriven sees the evaluation process as

a combination of a series of measures into a single quantitative or qualitative description of an entity. His formal definition of evaluation is:

> Evaluation itself is a methodological activity...The activity consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings, and in the justification of (a) data gathering instruments (b) the weighting and (c) the selection of goals."
>
> (Scriven, 1967,p.40)

The discrepancy between these two concepts of evaluation reflects the difference between what later became defined as Formative and Summative Evaluation. Formative Evaluation serves to improve a program and therefore deals with particular variables separately. Summative Evaluation serves to determine the overall merits of a particular entity, thereby allowing decisions to be made about adopting that program, continuing its operation, or alternatively rejecting it or discontinuing its operation.

The distinction between Formative and Summative evaluation first emerged in the evaluation of courses and instructional materials of different types. These terms were also later used in the context of learner's evaluation. Evaluation activities, which aimed to improve the learner's knowledge of what was taught in a course (and which consequently specified the achievement level of the individual in the full array of relevant educational objectives) became known as Formative Evaluation of Learners. On the other hand, activities which served to grade or certify a learner's results (expressed in terms of pass or fail marks) became known as Summative Evaluation of the Learner (Lewy 1985). The distinction between Formative and Summative evaluation became so fundamental to education, that shortly after the emergence of the concept "Self-Evaluation", one encounters the term Formative Self-Evaluation Studies. (see Duby 1986).

In addition to broadening the areas of educational enterprises, a differentiation occurred in the type of formative program evaluation studies. Cronbach's (1963) conception of Course Improvement referred to the empirical evaluation of draft curriculum materials. Very soon the term Formative Evaluation was applied to explorative studies carried out at the very early phases of program development, at a time when even the program draft was not available (Lewy 1977). Analytical studies, which did not use empirical data, such as Intrinsic

Evaluation Studies (Eraut 1985), and Curriculum Analysis Studies, became known as Formative Evaluation of educational programs. Moreover, the term Formative was Evaluation employed not only for studies aimed at improving the quality of course materials, but also for those that attempted to improve the diffusion and the implementation of programs (Tamir 1981).

A comprehensive study which summarizes theoretical and pragmatic aspects of Formative Evaluation was produced by Nathenson and Henderson (1980). This paper provides an orientation about some operational aspects of Formative Evaluation, and therefore describes methods and designs which have become widely used in evaluating educational programs.

The Link Between Program Development and Formative Evaluation

Formative evaluation helps program developers improve the quality of their products or their activities. It is a service-type activity, and is consequently expected to respond to questions, significant for the program developers themselves. In most cases the developers raise questions, and specify the nature of evidence they would like as a basis for answering their questions. The source of their questions may be differences of opinions among the members of the development team concerning a particular aspect or component of the program, or a checklist of formative evaluation questions routinely used by the development team. The topics for study are frequently suggested by the evaluation experts. They may call attention to some problems, or to alternative ways of action, which have been overlooked by the development team. As Alkin (1970) indicated in his frequently quoted definition of Decision Oriented Evaluation:

> "evaluation has been defined as the process of ascertaining the decision areas of concern, selecting appropriate information and collecting and analyzing information in order to report a summary of data, useful for decision makers in selecting among alternatives."
>
> (Alkin, 1970, p.2)

The fact that ascertaining the problem areas is listed as one of the components of the evaluation definition suggests that the evaluator, too, should assume responsibility for formulating questions. But no matter who generates

the evaluation questions, it is important that the evaluator and the development team agree about their importance and consider them a useful focus for an evaluation study. To decide whether a particular question constitutes a starting point for a study, the evaluator may employ the following criteria:

(1) The consequences of different responses. – The evaluator has to assess whether the response given to a particular question will affect the action of the development team.

(2) The feasibility of finding a decisive response to the question. – A question may reflect a serious dilemma, but it is advisable to use it as a starting point for a study, if there are reasonable chances that answers can be given to it on the basis of collecting and analyzing empirical data. The need to examine the feasibility of finding responses to an evaluation question has been pointed out in recommendations of the Joint Committee (1981). The committee suggests that the evaluability of a particular issue serves as an important criterion for deciding to launch a study.

(3) Priority of particular questions. – During the process of program production one may encounter many questions. While careful consideration should be given to all doubts and uncertainties encountered during such work, one usually has to limit the number of questions which will serve as a focus for an empirical study. Most of the questions have to be answered on the basis of deliberation among the parties involved. Decisions about what should be investigated through empirical studies can therefore be made only after reviewing all relevant questions, and determining their order of priority.

As already indicated, consensus among evaluators and their clients about the significance of questions, the need for empirical studies and the desired parameters of the study should be a prerequisite for launching a formative evaluation study. Figure 1.1 contains a schematic representation of the collaboration between clients and evaluators commonly experienced in Curriculum Development Centers or Institues (Lewy 1981).

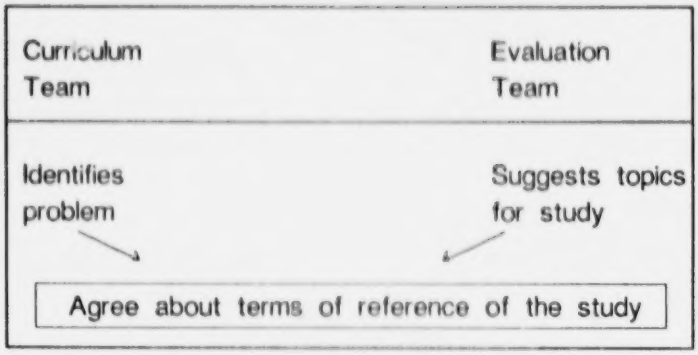


Figure 1.1 Starting a Formative Evaluation Study

An additional consideration in launching a Formative Evaluation study is the feasibility of providing evaluative results in due time. Program development itself requires a lot of time, and evaluation should not slow down this process. Evaluative information should be provided as soon as the program developers are to move to a new phase of activities.

The development team will be reluctant to wait for evaluation results, and if results are not presented at the critical turning point of the development process, one faces the danger of the evaluation findings being disregarded.

Evaluators should consider this and blame themselves if resarch findings are not used for the improvement of the program because of delays in presentation.

Program development can be divided into several phases. The questions faced by program developers vary from phase to phase. Evaluators have to be aware of the nature of questions which may emerge at various phases of program development and of the kind of actions which can be carried out at each of these phases. They should plan their evaluation activities accordingly.

The initial phases of program development are characterized by a relatively high level of openness, and program developers have great freedom to select from a large variety of alternatives. At the more advanced level of program development freedom of action becomes limited. The decision made at earlier phases operate as a sieve and block the free flow of divergent ideas.

Unfortunately, the working time allocated for various phases of program development is inversely related to the degree of freedom given to the developers. Usually, limited time is allocated for making decisions about the major parameters of their program. As the developers move towards more advanced phases of their work and enjoy a more relaxed time schedule, they have already lost a substantial proportion of their freedom of action. The relationship between time allocation and freedom of action during the process of program development is represented in figure 1.2.

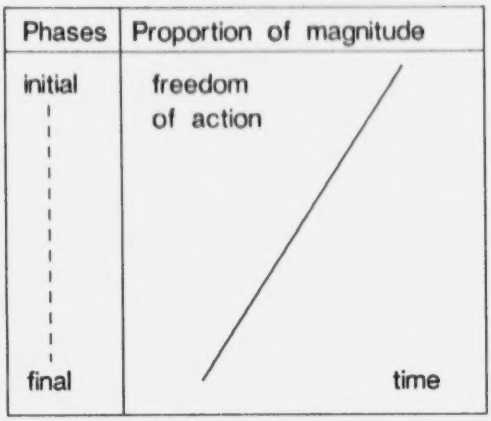


Figure 1.2 Time and freedom of action

Phases of Program Development

The time used for developing a particular educational program varies greatly from one project to another. The development of several complex innovative educational programs requires 3–4 years or even more. However, some are developed in much less time. But in most instances the time needed to develop an educational program is measured in terms of months and not in weeks or days. The development process consists of a series of sequential activities, which can be grouped into clearly defined phases. Sanders and Cunningham (1973) divided the development process into three phases: pre-development activities; the process of preparing the instructional materials (writing texts, creating teaching aids, working out blueprints for action etc.); and assembling the program components into a complete program kit.

Table 1.1 Phases of development and evaluation activities

Development	Evaluation
Pre-Development planning	Needs assessment Appeal studies Context studies
Production of prototype components and in-house assembly	Prototype evaluation Comparison of alternatives
Release of in-house version, preparation of revised edition	Intrinsic evaluation - checklists - criticism Empirical evaluation - tryout - error analysis - verification of structure
Publication of revised edition, diffusion and implementation	End of Formative Evaluation, start Implementation Evaluation and Quality Control Studies

In this article a modified version of this model is presented. A distinction is made between three program development phases: Pre-development planning activities; production of prototype program components and assembling the components; and the release of the in-house assembled program and preparation of the revised version of the program.

Table 1.1 lists the phases of program development and selected types of Formative Evaluation activities, which can be carried out at these phases. The scheme suggests that Formative Evaluation starts at the Pre-Development phase, and ends when the revised version of the program is published.

Pre-Development Phase

Program planning activities take place long before the team actually starts to prepare program elements or components. Indeed, the mere decision to produce a program constitutes part of the development activities.

Evaluation studies carried out at the **pre-development** phase frequently produce generalizations which touch on program development in general, and not only on the development of one particular program. In this article three types of pre-development evaluation activities are presented: Needs Assessment Studies, Appeal Studies, and Context Studies.

Needs Assessment Studies

Needs Assessment Studies have recently become a powerful instrument for improving educational planning. Educators frequently use the term **need** to imply a discrepancy between an existing situation and some desired set of circumstances. The term Needs Assessment is used for the systematic measurement of the discrepancy between the two (Pennington 1985). Suarez (1981) mentions that the term need is also used to refer to **preferences**, irrespective of whether a discrepancy is present. She quotes Scriven's definition of need, which says that needs signify a state in which a minimum satisfactory level has not been reached, or cannot be maintained. The reader may be aware that the first clause of Scriven's definition refers to the existence of discrepancy. The second clause broadens the definition of need and covers situations where the discrepancy is not yet present, but where it may appear without external intervention.

Needs Assessment Studies are used for various purposes in the context of educational planning. Suarez (1981) describes Needs Assessment Studies as defining educational goals. The studies also determine whether specific educational goals have been attained, and identify needs of students for instruction in a given subject area. They also identify needs of teachers for additional training, or future needs of an educational system in terms of resources, school buildings, manpower, instructional materials, etc. Due to the multiplicity of purposes, filled by Needs Assessment Studies in the process of educational planning, program developers should carefully select the study paradigm they intend to use. They should bear in mind that some widespread Needs Assess–

ment Study types have little potential to provide useful information for curriculum planners and for producers of new instructional materials. For example, a commonly-used Needs Assessment study paradigm is the administration of a standardized achievement test to the target population of learners. This identifies subject areas, or specific topics within the framework of a given subject in which the educational objectives have not been satisfactorily mastered by the learners. The findings of such studies may be useful for a variety of curricular decisions. It may, for example, result in recommendations, such as that time allocated for teaching a particular topic be increased; that a textbook used in school be replaced by another; that the teachers should develop supplementary exerises or instructional materials of other types for various topics taught in their classes, etc. The results of such a study can be interpreted as evidence of the need to develop new sets of textbooks, or instructional materials of other types.

In the context of curriculum development, Needs Assessment Studies are carried out in the initial phase of the development process. In most cases, institutes which assume responsibility for curriculum development allow relatively little time for conducting such studies. This is due to pressure exerted upon them to demonstrate their effectiveness by preparing and disseminating instructional products quickly.

Pennington (1985) suggests a series of information sources which may be used by curriculum developers to identify needs for certain types of instructional materials such as:

- analyzing relevant reports and publications
- drawing on the perceptions, expectations and experience of consultants
- reviewing existing curricular materials and critiques of their quality
- utilizing results of routine achievement surveys or examination projects
- comparing the characters of successful textbooks with those which are less successful
- organizing community forums to discuss needs for new programs

An example of using a large-scale community forum as a basis for determining needs in New Zealand is described in a government bulletin.

> "...to make the conferences more responsive to local views, the Minister invited six universities to plan and run seminars in their own regions...Two booklets, **Lets's Talk Education** and **Proposals for Change** were sent out. About 50,000 copies of each booklet were distributed free to all groups who, after discussion, recorded their views, opinions and recommendations, and forwarded them to the extension department. 8,000 submissions from 4,000 study groups representing 60,000 persons throughout the country arrived."
>
> New Zealand Educational Department Conference, 1974.

In large Curriculum Development Centers evaluation teams may gain time for carrying out comprehensive Needs Assessment Studies by creating an overlap between the initial development activities of a particular program and the overall assessment of needs. Since such centers usually prepare several sets of instructional materials simultaneously, it is possible to start development activities on the basis of well-established and broadly acknowledged needs. It is also possible to utilize the results of the Needs Assessment Study for balancing the areas of development activities, and if necessary, for adapting some products, not as yet fully developed, to the requirements suggested by the findings of the Needs Assessment Study.

This approach was used in the early years of the Israel Curriculum Center. Soon after the establishment of the Center several teams started to work on the development of new science courses for the school system. The existence of the need was established by drawing on the perceptions of experts. Parallel to the initial phases of the actual program development, however, the Curriculum Center launched a comprehensive Needs Assessment Study which attempted to derive educational objectives from the overall goal statements of the 1954 Compulsory Educational Act (Eden 1975). Within the framework of this study information was gathered from highly-renowned personalities such as representatives of the legislative body, leaders of the Teachers' Union, distinguished scholars in philosophy, education, sociology, psychology, economnics, etc.

Selected persons interviewed by the author of that study suggested operationally defined educational objectives which could be derived from the Compulsory Education Act. On the basis of these recommendations a list of ap-

proximately 200 operationally defined educational objectives was compiled. These goals were later rated by persons constituting representative samples of the following population groups: teachers, parents whose children studied in schools, educational administrators, and subject matter experts. They prioritized the objectives and assessed the discrepancy between the desired and the current status of these objectives in the educational system.
As soon as the results of the study were available, the curriculum teams checked their program, and if needed, introduced changes into it.

The results of a Needs Assessment Study have a time-bound validity. Needs change over time, and therefore institutes have to carry out such studies periodically and not only at the initial stages of their operation. The time pressure, which sets limits for the scope of Needs Assessment Studies in new institutes may decrease when the second and subsequent cycles of such studies are carried out in well-established institutes.

Studies of Appeal

It is well known, that children, and adults, learn better if they are interested in the topics being taught. Program planners therefore strive to find out what kind of programs are liked by the members of the target audience. They make an effort to devise learning activities which may capture their attention or stimulate interest. Of course in many curricula, and mainly in the subject-matter oriented ones, the major objectives and content areas of the studies are determined by the state of art of the subject itself. Even under such conditions, however, a great many options are open to the program developers, and by taking into consideration the preferences and dispositions of the learner, the success of the program may be greatly enhanced.

Even in school programs, where the learner is requested to participate in some presrcibed activity, interest remains an influential determinant of the quantity of learning. In formal learning situations the interest of the individual is a precondition of attending the program, let alone of benefitting from it.

A person who is not attracted by the title of an instructional-enrichment type television program, or who loses interest in a series of pictures which he/she incidentally watches, will quit the screen or switch to another program.

No wonder therefore that the issue of interest in, or appeal of, a program is a concern of producers of educational programs.

Appeal studies are classified as pre-development studies, because in most cases they have been carried out before program development has been started. In such studies the researchers draw conclusions from the current preoccupations of the target population, and make recommendations about issues related to the programs which will be produced. Appeal studies can also be carried out during later phases of the production, with the aim of drawing recommendations for revision of the program.

As indicated by Mielke (1983), the Children's Television Workshop (CTW) carried out a series of appeal studies.

Viewing preferences

The researchers wanted to know what interest children revealed in a variety of programs screened on television. The programs examined varied from each other in terms of format (e.g. magazine, drama), or genre (science, science fiction). They carried out a questionnaire survey in which the respondents had to mark their frequency of watching a variety of television programs. The results of this study provided orientation for the developers about the format and the genre preferred by members of the target audience.

Another technique used to determine viewing preference was a questionnaire attached to a simulated TV guide. This simulated guide specified parallel programs, and the respondent had to mark the particular program he would prefer to watch.

While these techniques of determining preferences were especially apt for planning programs for non-formal education, by analogy one may develop similar paradigms for determining the preferred styles of verbal presentation of learning materials in school books.

Topic preferences

Another question studied by the CTW evaluation team was the topic preference of children. This is equally significant in the context of developing instructional television courses and school textbooks. The children were presented with a large selection of photographs which covered a wide range of topics, such as animals, spaceships, plants, insects, etc. The responses requested were in the form of simple evaluative scales and forced-choice preference scales. Using a similar technique, Meyer (1973) asked seventh-grade children to make forced-choice selections from topics representing episodes related to the life of animals and the growth of plants. His study served to select topics for a course in biology. It is interesting to note that in both studies the children largely preferred animal topics. Mielke (1983) reported, that photos that involved animals in some emotionally engaging way were very popular.

Another technique for determining the appeal of various topics was eliciting questions from children, to which they wanted answers. The originality of questions asked by children exceeded the imagination of the adults. Some examples were:

> How does the body know when it is time to grow?
> Why do zebras have stripes?
> How do you make chalk?
> Is there a real Santa Claus?
> Why do I sometimes feel like I have no friends?

This is only a very small selection of questions. Each of them could serve as a topic for a graduate course in various disciplinary or interdisciplinary areas.

Context Evaluation

No curriculum operates in a vaccum. A series of instructional materials that can be successfully used in one context, may not be appropriate for use in another.

Curriculum developers have to obtain specific information about various aspects of the context that will affect the utilization of a new educational program. Stufflebeam et al (1971) in the well-known CIPP (Context, Input, Process,

Product) model of Curriculum Evaluation lists context evaluation as the first task area in a sequence of four evaluation task areas. The four letters of the CIPP acronym indicate that in each particular study one has to consider four significant areas of the entity being evaluated. By no means do they suggest sequential steps of evaluation foci. Nevertheless, in practice it transpires that one may start conducting studies about the context of a planned program prior to the specification of its exact parameters. Indeed, Stake (1967) uses the term **antecedents** to denote activities which partly overlap with what the CIPP model designates as Context evaluation. The semantic meanings of the two terms suggest that Stake's model concerns itself with the time sequence of evaluation activities, while the CIPP model put emphasis on the multiplicity of program aspects which have to be examined. Still, both modes identified areas of concern which can already be examined at the pre-development phase.

Stufflebeam (1971) claims that for the purpose of context evaluation information must be obtained about **needs** which have been met, unused **opportunities** and **problems**. Thus, he includes Needs Assessment Studies under the broader heading of Context evaluation. In practice, however, context evaluations deal with a broad variety of phenomena which traverse the boundaries of Needs Assessment Studies. Examples are:

- What are the implications, if any, of legislative acts on the school curriculum?
- Identification of court decisions, which may have implications for school programs.
- The existence of an appropriate network for disseminating new programs.
- The availability of equipment for implementing programs of different types, or the possibility of acquiring such equipment.
- Parameters of school climate, such as teacher-pupil relationship, definition of the teacher's role, and the freedom of the school and teacher to determine its curriculum.
- Community interest, the concern of religious institutions, voluntary organizations' initiatives, parents' involvement in educational programs.
- Economic trends and trends in the occupational structure.
- Demographic trends: changes in birth rate, internal immigration.

- The attention paid by the mass media to matters of education and its implications.
- Differential needs of various geographical, ethnic and linguistic groups in the country.
- The general level of educational achievement in the whole educational system compared with others.
- Changes in overall achievement level: decline or growth.

These items constitute a sample of issues which define the context of an educational program. It is clear that all these issues cannot be investigated as a preparatory activity for any single educational program. Evaluators have to locate available information and determine research priorities in order to establish a working program, which can be executed in the face of constraints such as resources, time, and manpower.

Production of Prototype Components and Their Assembly

The Curriculum reform of the 60s and 70s brought forward the demand for the empirical tryout of instructional materials prior to their widespread use. The high cost of producing new instructional materials was one of the factors that precipitated the demand for empirical evidence about the quality of new programs. Apart from such demands, however, curriculum experts strove to realize Campbell's (1969) vision of the experimental society, where the preliminary tryout of social action is an adjunct to planning.

Evaluators believe that there is no need to postpone evaluation activities till program developers prepare a whole course of study. During the process of the program development a particular component of the program can be evaluated as soon as it is available. In this section two Formative Evaluation designs are described: Prototype Evaluation and Comparison of Alternatives.

Prototype Evaluation

The tryout of instructional materials is usually carried out in successive phases. Individual units of a program can be empirically tried out prior to their assembly into a complete program kit.

The curriculum team may invite individual students and examine their reaction to some typical components or learning units of the new program. One should be aware of the artificiality of such situations. They do not fully reproduce the routine situation of using curriculum materials in the classroom, but even under such artificial conditions one may gain useful insights into the merits and flaws of the program.

The major advantage of evaluating a program unit immediately after its production is the possibility of revising it before starting work on the production of subsequent units. This may eliminate an accumulation of flaws, and substantially reduce the number of iterations needed for completing an acceptable version of the full program.

Such a piecemeal evaluation of separate units poses serious logistic al problems. The evaluator may invite individual students to evaluate the first learning unit of a program, but he or she cannot then move smoothly to the evaluation of the subsequent units, which have not yet been produced.

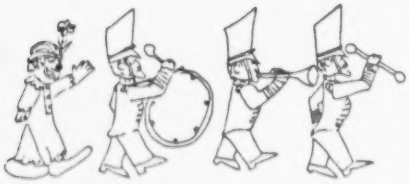
The production of the subsequent unit may require a long time, and by the time it becomes available for tryout those who participated in the tryout of the first unit may have forgotten what they have learnt. The evaluator faces reteaching the first unit to those who participated in the earlier tryout, and then teaching the new unit. Alternatively, he could invite a second group of individuals and teach the two units one after the other.

Of course such constraints emerge only if the units have a hierarchical structure, and mastery of a certain unit is a prerequisite for studying the subsequent one. But a great many school programs have six to eight hierarchically structured units, and consequently their piecemeal tryout may reqire a complex pattern of iterative teaching for evaluation.

This piecemeal tryout of instructional materials is usually carried out by using "soft" methodologies. The curriculum team, together with the evaluator, observes the learner's behaviour, takes note of the learner's responses to the instructional materials, and if needed, suggests specific modifications. The evaluator may be compelled to use non-representative samples of the target population. Evidence has been accumulated that even such limited tryouts

may contribute to the improvement of the program (Markele 1967). The major methodological problem at this phase is usually referred to as a **distortion filter**. The evaluator does more than note what he or she observes: he or she interprets or "filters" what is observed. The ideal tester of instructional materials should be sensitive to subtle manifestations of the pupils' problems.

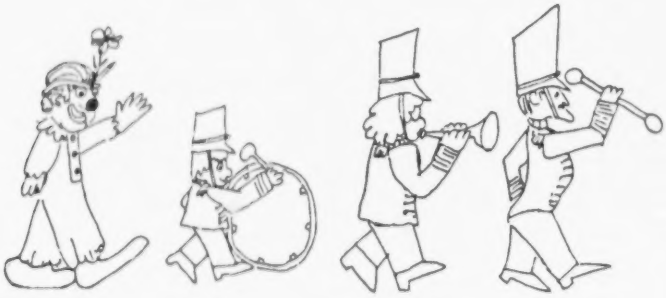
An instance of trying out a curriculum component and revising it on the basis of the tryout result is reported in EPIE (1980).



Original Directions

Underline the man behind the man with the drum.
Circle the man in front of the man with the horn.

Revised Directions



Look below at the people who play music in the parade. Point to the lady with the horn. Now point to the one in front of the lady.

Point to the boy playing the drum. Now point to the one behind the boy. Underline the one behind the boy.

Figure 1.3 An example of a picture and its revision

Figure 1.3 contains the original version of an illustration appearing in a primary school book and the exercise related to it.

The proportion of the correct responses to the revised exercise and illustration was substantially higher than to the original.

Comparison of Alternative Versions

Evaluation reports contain descriptions of experimental studies which were carried out to determine the relative effectiveness of alternative program components or units. Comparison of alternative treatments, or more frequently comparison of an innovative treatment with the traditional one represents the classical design of evaluation, and numerous experts still strongly rcommend this design for proving the worth of new programs (Rieken and Boruch 1974). Curriculum Evaluation has broadened the scope of legitimate evaluation designs, and also proved the viability of non-comparative studies. Nevertheless, whenever the utilization of the experimental design is feasible, it is wise to seize the opportunity.

Preparing alternative versions of a whole course would require prolonged development activities and great financial investment. Even if funds were allocated for such development activities, however, one could hardly derive useful generalizations from such comparisons. It would be difficult to determine what features of a particular program were superior over others. It is much more feasible to compare various versions of a short curricular unit. It does not require extensive development activities, and the comaparison of the alternative versions can be carried out during a relatively short time. Most importantly, it may be easier to create a single conceptually definable variation among two short curricular units than between two whole courses. One may say that short units are less vulnerable to contamination variables than longer ones.

Indeed, most studies which employed the experimental design used short teaching units, and differed from each other with respect to an educationally significant variable. The findings of such studies served as a basis for generalizations about preferred curricular features in general.

The examples presented here touch upon issues of genre, and transmission technique.

Differences in genre: descriptive-scientific or narrative prose

A biology curriculum for 12–year-olds dealt with the topic of **animals and their Habitat**. Two short units of the course were developed in two different genres of prose: descriptive scientific prose and narrative prose. The descriptive-scientific genre followed the traditional science textbook style, and the narrative genre presented the scientific facts and priciples interwoven with story adventures. An experiment was carried out in which one group of classes studied the descriptive scientific version of unit 1, and the narrative version of unit 2. The other group studied the alternative version of the two units. At the end of the experimental teaching the attitudes and cognitive achievements of pupils in the two groups were examined. The narrative genre was found to yield better results on both variables and for both units (Rosenblum 1973).

Differences in teachers' behavior: signalling vs. non-signalling

The DISTAR curriculum packages published by Science Research Associates represent the method of direct instruction (Rosenshine 1985). Each lesson is a complete script, which the teacher follows verbatim to present the instructional task-cue for children to respond to, and to provide corrective and reinforcing feedback (EPIE 1980). The program developers conceptualized the question and response sequences as a stimulus response chain. Nevertheless, some "slow" children could not react to the teacher's question, but instead joined in the "choir", and recited the responses uttered by those classmates who responded quickly.

To create a smoother flow of stimulus-response exchange, the developers prescribed that the teacher should give a signal, similar to the signal given by the conductor to the orchestra. The children were asked to give their response when the signal was given.

To examine the effectiveness of the signal a comparison was made between the proportion of delayed responses observed under signalling and non-signalling conditions. The analysis revealed that the signals reduced the delayed responses from about 5% to 1%.

The findings reported in this study, and in previous studies, provided generalizations which led to the structuring of the whole program, and not only those units subjected to evaluation.

Release of the In-House Assembled Version

With the release of the in-house assembled version of a program it is possible to carry out, a full-fledged trial of the whole course. At the same time the development team may start to work on the revision of the program and make the necessary steps for publishing a revised version.

The evaluation activities carried out at this phase are dealt with in two sections: intrinsic evaluation, and curriculum tryout.

Intrinsic Evaluation

In this context, the adjective **intrinsic**, refers to evaluation studies which are not based on empirical data. As Scriven (1967) indicated, some aspects of educational programs, such as the legitimacy of the program goals, cannot be evaluated on the basis of empirical data only. The approval or rejection of a particular education goal may reflect value positions of interested parties.

Curriculum components may be already subjected to intrinsic evaluation at the preassembly phase of program development. Thus, for example, curricular objectives can be evaluated immediately after they have been formulated, and long before the full completion of the program development activities. Nevertheless, intrinsic evaluation frequently addresses some overall aspects of an educational program, which can be examined only when the whole program package is available. For example, the internal consistency of a program or the relatedness of the learning activities to stated program goals can be examined only by inspecting a fully-assembled program kit (Eraut 1985).

Evaluation checklists

Checklists to evaluate the quality of instuctional materials were used long before the term Formative Evaluation was coined. Devereux (1933) summarized attempts to use checklists for the evaluation of educational "talking pictures" during the early decades of this century. Eraut (1985) notes that there is a continuous growth in the quantity of checklists used for evaluation educational

programs. He lists four reasons for this. First, conditions for developing educational programs differ from each other and a checklist which fits the evaluation of one program may not be appropriate for another. Secondly, there is no unanimous acceptance by experts of a single learning theory. Thirdly, the context of using a particular program changes from case to case. Finally, learners differ from each other. Nevertheless, numerous curriculum checklists are stated in quite general terms and are rcommended by their authors for general use of program evaluation. Bloom (1977) constructed an evaluation checklist of ten items which can be examined on the basis of judgemental evidence. The items and the type of evidence required are reproduced in Table 1.2.

Table 1.2 A ten items checklist

	Judgmental Evidence					
Issues Being Considered	Subject Expert	Other Specialists	Teachers and Supervisors	Students	Observations in the Classroom	Student Learning
A. Accuracy and soundness of materials	X	X	X			
B. Internal structure and sequence	X	X	X			Criterion-referenced formative test
C. Feasibility, cost and practicality	X	X	X		Problems posed for teacher training	
D. Clarity and meaningfulness	X	X	X	X		What has been learned well?
E. Level of dificulty	X	X	X	X	Student responses	Overall effectiveness
F. Interest value for students	X	X	X	X	Student attention and involvement	Interest and attitude questionnaires and scales
G. Special teaching problems	X	X	X		Condition of instruction	What needs to be revised?
H. Relevance to students' needs and aspirations	X	X	X	X		
I. Overall reactions to materials	X	X	X			
J. Unintended learning and effects	X	X	X		General and specific comments by observers	

As specified in Table 1.2, Bloom suggests combining judgement data with empirical data. Nevertheless, evaluative statements about each of the 10 items listed in his model can be derived even if only judgemental data are available.

The most widely-known program evaluation checklist was produced by Eash (Walberg 1974).

The Eashs checklist has been reproduced in several evaluation textbooks, and hence its popularity.

Table 1.3 Excerpts from Eash's model

I.	**OBJECTIVES** What is the nature of the general goals of the material stated? Are specific objectives stated for teacher use?
II.	**ORGANIZATION OF MATERIALS (SCOPE AND SEQUENCE)** What is the scope of content covered in the matateriels? How is the scope of the materials organized?
III.	**METHODS OF INSTRUCTION** What role is emphasized in the method : teacher, pupil, or both? What are the specific features of the method or methods recommended? Are the specific features of the method such that require the teacher to do extensive prior preparation or participate in specific training?
IV.	**EVALUATION** What test materials are included for the students' and teachers' use? Are the test items adeqate for informing a teacher of the students' progress toward the instructional objectives set for the materials? What do the tests measure?

The Eash's model adopted the Tylerian approach to curriculum theory and the items contained in it are grouped into four clusters: objectives, organization of materials, methods of instruction, and evaluation methods for examining outcomes.

Finally, a checklist developed by Gall (1981) represents an eclectic approach. Items contained in it are grouped in four clusters. Illustrative items from this are presented in Table 1.4.

Curriculum criticism

Curriculum criticism as a type of curriculum evaluation activity was first advocated by Mann (1969), and became later developed and expanded into an operationally powerful way of improving educational programs by Eisner (1979). The idea of employing this method as a useful tool for curriculum evaluation has its roots in the awareness that instructional materials are cognate to literature. Both contain verbal materials, moreover, usually both are published in the form of books or articles. Drawing an analogy between the evaluation of literary objects and instructional materials represents an attempt in the educational literature to pay attention to evaluative methods employed in other disciplinary areas. It opened the road for utilizing evaluation experience accumulated in other areas, like investigative journalism, law, photography, ethnography, etc., for program evaluation.

The affinity of the evaluative work in the field of literature and curriculum materials is noted both by curriculum evaluators and literary critics.

Willis (1978) stated that curriculum

> "...is an art as well as a science. While it is well for the curriculum worker to know the scientific principles, the empirical data, and the technological applications, that inform much of his practice, it is equally well for him to be conversant with those aspects of arts which come into play, when what happpens with a curriculum or in a classroom inevitably spills out beyond those scientific constucts he has chosen to employ."
>
> (p. 93)

Table 1.4 Excerpts from an eclectic checklist

PUBLICATION AND COST

Authors: Are the authors known and respected professionally?

Cost: Is the cost of the materials reasonable relative to other comparable materials?

Development History: Were the materials adequately field tested and revised prior to publication?

PHYSICAL PROPERTIES

Aesthetic Appeal: Are the materials likely to appeal to the user's aesthetic sense?

Components: Do the materials contain so many components that the teacher will have difficulty in keeping track of them and using them

Consumables: Does the product make unnecessary use of consumable materials?

CONTENT

Approach: Does the developer use an approach consistent with the district's curriculum?

Instructional objectives: Are the materials' objective compatible with the disrict's curriculum and acceptable to teachers?

Instructional objective-types: Do the meterials contain affective objectives in addition to cognitive objectives?

INTRUCTIONAL PROPERTIES

Assessment Devices: Do the materials contain tests and other assessment devices that will be helpful to the teacher?

Comprehensibility: Will the materials clearly be understood by the students who will be using them?

Coordination with Curriculum: Are the materials compatible with other materials currently being used in school?

Similarly Hirsch (1976), whose writings are well known in literary criticism, emphasized the common nature of those cultural objects, which are based on the utilization of the language.

> "A technical essay, an ordinary conversation or a poem has, therefore, necessary, inherent values: of course the values are different, but the structure of arguments for their existence is the same. It follows that there is no reason for isolating literature and art in mysterious ontic realms apart from other culutral realities."
>
> (p. 109)

The affinity of a literary object with curriculum materials is described by Mann (1969), who claims that both select events from an undifferentiated totality of life. The author chooses one feature of life unstated. Acting in this way the author creates a shape and meaning for life. The curriculum developer, too, selects phenomena from the world or reality and represents them in a textbook or in a set of instructional materials.

The aim of the literary critic, as well as that of the curriculum critic, is to disclose the meaning of the object being criticized, or to illuminate its qualities. In this view, the critics task is not primarily to issue a judgement, but rather to lift the veils that keep the eyes from seeing (Dewy 1934).

Eisner (1979) makes a distinction between connoisseurship and criticism, defining the first term as the art of appreciation, and the second as the art of disclosure.

> "The conoisseur recognizes the qualities of a particular object. Connoisseurship is a private act; it consists of recognizing and appreciating the qualities of a particular object, but it does not require either a public judgement or a public description of those qualities."
>
> (Eisner 1979, p. 193)

To develop connoisseurship one must accumulate experience. For example, the wine taster must drink a lot of wines and learn how to recognise their qualities. The connoisseur of instructional materials has to learn to perceive the subtle particulars of the educational life, and their role in the classroom. Criticism is the public side of connoisseurship. The critic must be able to specify his judgement, to explain, and to justify it.

It seems difficult if not impossible to provide operational guidelines for producing curriculum critiques. Moreover, literary experts claim that one cannot specify a set of criteria which should be considered by the critic. Harrison (1968) claims that the art critic does not operate with a closed set of criteria. The definition of criteria constitutes a part of the process of criticism. The critic is the source of new ideas. The same can be said with regard to curriculum criticism. The curriculum critic does not operate according to a preordinate pattern of enquiry. He or she does not start by specifying a closed set of criteria. The starting point for the criticism is the text, the instructional materials or the actual classroom situation. They constitute the basis for deriving criteria.

Curriculum criticism may fulfil both formative and summative evaluation roles. The provision of information about the traits of the curriculum, its hidden meaning, the client's value orientation, ethical conceptions, and pedagogical principles represented in the program may help people to make decisions about using a particular program. In this respect its role is summative.

Disclosure of the meaning of a program as perceived by critics may also help the developers obtain a clear conception of what their program represents in reality. Criticism may raise second thoughts in the mind of the developers about the qualities of the program and its merits. It may strengthen satisfaction with their work, or it may reveal flaws and motivate the developers to modify the program. In this respect literary criticism is different from curriculum criticism. Literary criticism addresses a completed object of art. It discloses its meaning, but seldom motivates the author to introduce changes in the object of the criticism. As a matter of fact, art criticism deals with unchangeable objects. In contrast, curriculum criticism, though it deals with finished products or objects, they never reach such a level of completeness that their elements cannot be changed. Unlike literature, the new editions of curriculum packages contain modifications, which frequently reflect reactions to curriculum criticism. Hence the formative role of curriculum criticism.

Curriculum Tryout

Commercially distributed curriculum packages usually cover the content of at least a one-year course in a given subject. Still, schools prefer to use series, which provide instructional materials for several consecutive grade levels. Across the whole world it is common to find series of textbooks in mathematics, science, mother tongue, etc., for all primary, or middle grades, and quite frequently for all grades. The designation K–12, which is attached to textbooks available for all grade levels from kindergarten to grade 12, is a highly attractive recommendation for schools to adopt a certain textbook.

From an organizational and pedagogical point of view, a school will prefer to adopt a textbook if it covers the content of at least a whole year. This scope of coverage at least is needed to try out new curricular materials in a natural environment, and under routine conditions. The tryout of a shorter curricular unit, covering less than a year's content of a given subject, may interfere with the routine work of the class. For this reason Curriculum Development Centers and Institutes carry out tryial studies of newly developed materials as soon as they have completed the assembly of curriculum kits, covering the content of the syllabus for at least a year. The tryout is usually carried out in a few classes only. Experts recommend testing new materials in a group of 6–10 classes, selected on the basis of judemental sampling princples (Lewy 1977).

The small-scale tryout of instructional materials creates a problem with regard to the supply of these materials. The quantity of books or other types of materials needed for the tryout is very limited, and it may be difficult to find commmercial companies who will assume responsibility for producing the needed materials.

Moreover, the preparation of such editions of a curricular set may require a considerable amount of time, and would prolong the already "unbearably" long period of program production.

To cope with this difficulty, an interim edition of the curricular materials is often produced, in a much simpler form than the final product. The verbal matnerials may appear in mimeographed pages. Very few colour pictures, if any, will be used in this edition, and the assembly of the materials may be done by in-

house teams. In this respect, the tryout situation will be quite different from the eventual open utilization of commercially produced, later editions. For practical reasons, however, one has to cope with this problem of artificiality.

Since the tryout is usually carried out with instructional materials of lower quality than those used at a later phase, one should be concerned with the predictive validity of tryout studies. Unfortunately, little attention has been paid to this problem, and no research evidence is available about its effects. In instructional film development, attempts were made to compare the learning outcomes of the final film product with those obtained from using the preliminary picture-board version (a series of still pictures drawn as a blue-print for carrying out the shooting). While the final version produced a better outcome than the preliminary picture board version, it was proved that the picture board version adequately predicted the profile of outcome. At the program segments where the story board produced a relatively low outcome, the corresponding film segment, too, produced a relatively low outcome (Zuckerman 1954). By analogy, one may assume that the home assembly version of textbooks is less effective than textbooks of higher technical quality. Nevertheless, their predictive validity renders the tryout a useful diagnostic device for identifying program flaws.

The design of tryout

The tryout of curriculum materials, in most cases, represents a one-shot non-experimental study. Criteria for examining the success of the program are predetermined, and consist of process and product variables. Examples of these are: interest revealed in the studies, participating in learning activities, initiative on behalf of the learners, friendly school climate, development of desired attitudes, internalization of values, acquisition of knowledge, and mastering of educational objectives. Standards of achievements are determined in criterion-referenced style, and neither pre-test measures nor comparisons with control group achievements are needed.

The futility of comparing outcomes of various curricula was convincingly demonstrated by Walker and Schafferzick (1974). Different curricula have different objectives, and the tests used to measure end-of-course achievement are mostly biased towards one or another set of instructional materials.

Curriculum developers, therefore, usually specify expectations in terms of "X percent of learners attaining Y percent of curricular objectives". Some curriculum groups put the figure 90 in the place of both X and Y. Others satisfy themselves with more modest expectations, and use figures as low as 70.

Curriculum tryout deviates from the classical experimental design in another respect, too. Experimental design requires strictly defined standard treatment across all subjects. The experimenter is supposed to be a neutral observer, and is not allowed to intervene in the process of treatment. In the case of curriculum tryout the experimenter is encouraged to intervene and change the instructional materials whenever it seems necessary.

Error analysis

As indicated in the previous section, curriculum tryout is designed to identify weak spots in the instructional materials, and thus to enable improvment of the revised version of the program. A direct method of doing this is Error Analysis.

Error Analysis in the context of curriculum evaluation is defined as the analysis of students' work – home work, school work, and tests. It identifies major errors in the students' work and any mislearning, which might occur (Postlethwaite 1985). Postlethwaite specifies three ways of carrying out error analysis for formative evaluation.

One may first identify objectives, which have not been mastered by students. As a hypothetical example he presents the data reproduced here in Table 1.5.

Table 1.5 Percentage of correct responses in a sample of six schools.

	Schools						
	A	B	C	D	E	F	Average
Item 1	90	80	85	87	93	78	85
Item 2	20	30	27	19	15	22	22
Item 3	90	80	15	20	50	90	57

The data presented in Table 1.5 indicate that the objective represented by Item 2 has not been mastered by the learners. Though the data presented here do not reveal the reason for failing to master that particular objective, a thorough analysis of the question and of the textbook may provide useful clues as to the reason for the failure.

A second approach to error analysis is to draw profiles for the achievement of individual learners in classes and to compare this with information provided by the teacher about the objectives the learners were expected to master. This method takes into account the differential emphasis of various objectives across classes or between individuals.

Finally, Postlethwaite (1985) describes a mode of error analysis, in which the distractors used in a multiple choice test contained hints about the nature of the error committed by the learner.

This is explained through the example:

$$(-17) + (-14) =$$

(a)−31 (b)+31 (c)−3 (d)+3

If the student answers (b), the most likely error is a result of confusing

$$(-)+(-) \text{ with } (-)\times(-)$$

Light and Reynolds (1972) described a more complete design of Error Analysis, which also comprised the process of debugging program errors. They specified three steps of this process. First, the evaluator had to identify a case where a student did not do well in a test. Secondly, he had to identify the cause of the problem. Thirdly, it was necessary to find an effective solution to the problem.

A successful way of identifying the cause of the error was to observe similarities between the items missed in the test, how they differed from those passed in the test, where the relevant content was presented in the instructional materials, what aspect of the instructional materials could have caused the failure, and finally how the hypothesized cause of failure could be tested. The authors provided several examples of using this sequence of questions for identifying causes of error. They felt that formative evaluation was complete only if the findings of the evaluation actually led to the improvement of the program. They therefore greatly emphasized the need for identifying causes of an observed error.

Verifying a hypothesized structure

The topics addressed within the framework of a course are related to each other according to a predetermined pattern of organization, constituting a structure. Curriculum structures reflect certain views about how learning should take place. Contemporary course materials differ from each other in regard to their structure. Some of them have a modular structure, i.e. the teacher is expected to determine the sequence of teaching various curriculum units in the class. Other curricula exhibit a "discovery learning" structure, i.e. learning experiences are sequenced in such a way, that a series of concrete facts or occurrences enable the learner to discover an explanatory principle or generalization.

Curricula may have a predetermined hierarchical structure, where the mastery of some units or sub-units are a prerequisite for mastering others. Gagne (1970) espoused the utilization of hierarchical structure in course materials, and provided theoretical and empirical evidence about the effectiveness of this structure. Bruner (1960), on the other hand, preferred the discovery method.

The organizational structures listed above are examples of a broad inventory of principles used for determining the sequence of activities in a series of instructional materials in the class. Most frequenly, course developers explicity state the organizationl principles used in their programs. It is not enough to indicate this however. One has to provide evidence that the statement about the structure is not only a declarative one, but that it reflects a certain reality.

As structures differ from each other, so the designs used for the verification of their existence also differ from each other. A design used to verify the modularity of curricular units is quite likely in adequate for verifying the existence of another organizational structure.

For the sake of illustration, details of a formative study used to verify the existence of a hierarchical structure in a primary school arithmetic program are presented here (White and Gagne 1978).

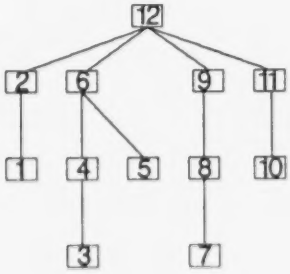


Figure 1.4 A hierarchical structure of a learning unit.

The unit contained twelve sub-units. The developers of the program hypothesized the existence of a hierarchical relationship among the twelve sub-units. The hypothesized learning hierarchy of the units is illustrated in figure 1.4. According to the scheme, sub-unit 1 is a prerequisite for sub-unit 2, and sub-unit 2 is prerequisite for sub-unit 12.

White's (1974) method for verifying learning hierarchies was employed in the study. According to this method one has to construct test items which examine the mastery of each particular sub-unit. One then has to examine the

response pattern of each student to each pair of items. The four possible patterns of responses can be represented in the following way:

(1) + + i.e. both responses are correct

(2) – – i.e. both responses are incorrect

(3) – + i.e. the first item incorrect, the second correct

(4) + – i.e. the first item correct, the second incorrect

If pairs of items are sequenced in such a way that the first one represents a lower hierarchical position, and the second a higher level, then the pattern (3) response invalidates the hypothesized relationship. The actual results obtained are presented in Table 1.6 .

Table 1.6 Pass and fail for hierarchy pairs of items

		Pass (+) and fail (–) categories in the lower–higher			
		1	2	3	4
Relation	White's finding	(++)	(––)	(–+)	(+–)
1 – 2	Accept	123	19	1	5
3 – 4	Accept	105	21	0	22
4 – 6	Accept	102	40	3	3
5 – 6	Accept	105	24	0	19
7 – 8	Reject	62	36	33	17
8 – 9	Reject	24	42	71	11
10 – 11	Accept	79	49	1	19
2 – 12	Accept	44	24	0	80
6 – 12	Accept	40	39	4	65
9 – 12	Reject	13	82	31	22
11 – 12	Accept	42	66	2	38

The results provided a partial verification of the hypothesized structure. Among the relationships examined, eight hypothesized relationships were proved correct, and three incorrect.

Findings of this type have practical value for the revision of the program. The subunits which do not fit into the hypothesized hierarchical structure either can be deleted from the unit, or its location in the sequencing of activities should be other than the actual one.

The Revision of the Curriculum

A model representing the process of curriculum revision is presented in Figure 1.5 .

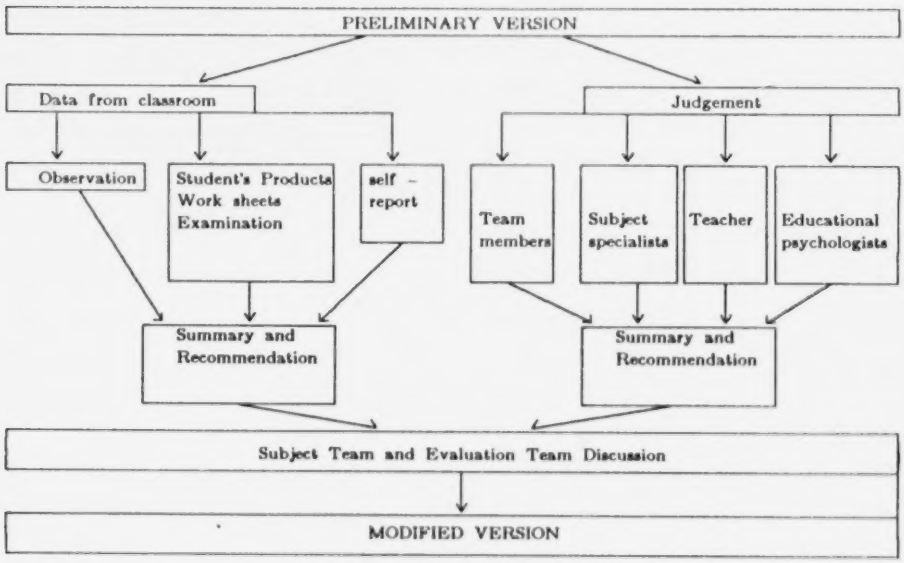


Figure 1.5 A model of curriculum revision*

* The model was developed by M. Silberstein at the Israel Curriculum Centre.

It can be seen that information gathered from various sources is integrated and constitutes input for decision making.

Concluding Remarks

Despite the tremendous growth in the number of formative and summative evaluation studies published in the 1960s and 1970s, theoreticians have not produced rules as to the procedures appropriate for either type of study. Scriven plays down the differences betwen formative and summative studies. Nevertheless, several reviews of empirical studies have pointed out systematic differences.

Stake (1977) provides a series of terms which characterize differences in information sought by users of formative and summative evaluation. Formative evaluation focuses on relatively molecular analyses, such as cause seeking, and is interested in the broader experiences of the program users. It tends to ignore the local effect of a particular program. Summative evaluation focuses on molar analyses, provides descriptive information, is interested in efficiency statements, and tends to emphasize local effects.

Alkin (1974), analyzed 42 evaluation reports and questionnaires completed by project directors, and listed additional characteristics of formative evaluation studies. Their design is characterized as exploratory, flexible, focusing on individual components of the program, and as emphasizing iterative processes. While not comparative, it seeks to identify influential variables. A formative evaluation study uses a great variety of instruments which are either locally developed or standardized; it relies on observation and informal data collection devices, most of which are locally chosen. In contrast, summative evaluation studies tend to use well-defined evaluation designs, which are as unobtrusive and non-reactive as possible; they are comparative and concerned with a broad range of issues – – for example, implications, politics, costs and competing options. The instruments used in summative evaluation are publicly accepted, reliable, and valid instruments, and reflect concerns of the sponsor and the decision-maker.

The formative-summative distinction has increased the range of evaluation studies and contributed to the improvement of educational planning. Three decades of utilizing these terms has produced little consensus concerning their distinct features, however. The allegation that they differ only from the point of

view of the decisions they are supposed to support, and that no distinction should be made between them from the point of view of design, methodology, etc. has been endorsed by many evaluators without its veracity having been empirically examined. At the same time, other evaluators have asserted that methodological rigor is only required in summative studies, while in formative studies the evaluator may rely on less rigorously validated data and analysis procedures.

This claim has never been confirmed by evaluation experts, but has nevertheless led to the burgeoning of poorly-designed formative evaluation studies.

To avoid such pitfalls, it is necessary to conduct empirical studies to ascertain the characteristic method and design features of formative and summative evaluation. Such studies might facilitate the use of formative and summative evaluation, and enhance their quality.

References

Alkin, M.C., 1970, Products for Improving Educational Evaluation. **Evaluation Comment**, 2(3), 1–15

Alkin, M.C., 1974, **Evaluation and Decision Making** Los Angeles, University of California, Center for Study of Evaluation

Bloom, B.S., 1977, Tryout and Revision of Educational Materials and Methods. In A. Lewy (ed.) **Handbook of Curriculum Evaluation**. New York: Longman

Bruner, J., 1960, **The Process of Education**. Cambridge, MA: Harvard University Press.

Campbell, D.Y., 1969, Reforms as Experiments. **American Psychologist**, 24, 409–29.

Cronbach, L., 1963, Course Improvement through Evaluation. **Teachers College Record**, 64, 672–683.

Devereaux, F.L, 1933, **The Educational Talking Picture**. Chicago: University of Chicago Press.

Dewy, J., 1934, **Art as Experience**. New York: Minton, Balch and Co.

Duby, A., 1985, Self Formative Evaluation of Instructional Materials with Reference to Games and Television. Doctor's Thesis submitted to University of South Africa.

Eden, S., 1975, The Translation of General Aims into Functional Objectives: A Needs Assessment Study. **Studies in Educational Evaluation**, 1, 5–12.

Eisner, E., 1979, **The Imaginative View**. New York, MacMillan.

EPIE, 1980, Educational Products Information Exchange. **Deciphering LRV**, vol 12(3).

Eraut, M., 1985. Intrinsic Evaluation. In Husen T.and Postlethwaite T.N.(eds.) **International Encyclopedia of Education.**

Gagne, R.M., 1970, **The Conditions of Learning** (2nd ed.) New York: Holt, Rinehart and Winston.

Gall, M.D., 1981, **Handbook for Evaluating and Selecting Curriculum Materials.** Newton: Allyn and Bacon Inc.

Harrison, B., 1968, Some Uses of Good Criticism.
In Francis J. Coleman (ed.) **Aesthetics**. New York: McGraw Hill.

Hirsch, E.D., 1976, **The Aims of Interpretation**. Chicago: University of Chicago Press.

Husen, T., and Postlethwaite, T.N. (eds.), 1985, **International Encyclopedia of Education: Research and Studies**. Oxford: Pergamon Press.

Joint Committee on Standards for Educational Evaluation, 1981, **Standards for Evaluation of Educational Programs**. New York: McGraw Hill.

Lewy, A. (ed.), 1977, **Handbook of Curriculum Evaluation**. New York: Longman.

Lewy, A., 1981, Sharing Responsibility for Evaluation. In Lewy, A. et. al., **Decision Oriented Evaluation: The Case of Israel.** Rechovot: Balaban Publ.

Lewy, A., 1985, Formative and Summative Evaluation. In Husen T. and Postlethwaite T.N. (eds.) **International Encyclopedia of Education.**

Light, J.A., and Reynolds, L.G., 1972., Debugging Product and Testing Errors. **Viewpoint,** 48, (4), 45–78.

Mann, J.S., 1969, Curriculum Criticism. **Curriculum Theory Network** issue 2 Winter p. 2–14, Also in Willis G. (ed.) **Qualitative Evaluation.**

Markele, S.M., 1967, Empirical Testing of Programs. **Sixty-Sixth Yearbook of the National Society for the Study of Education,** part II. Chicago: University of Chicago Press.

Meyer, M., 1973, Differential Attitude of Pupils and Teachers to Various Topics in Science Studies (Hebrew). Jerusalem: Hebrew University, Master's Thesis.

Mielke, K.W., 1983, Formative Research on Appeal and Comprehension in 3–2–1 Contact. In Bryant J. and Andersen D.R. (eds.), **Children's Understanding of Television**. New York: Academic Press.

Nathenson, M.B. and Henderson, E.S., 1980, **Using Student Freedback to Improve Learning Materials.** London: Croom Helm.

New Zealand Education Development Conference, 1974, **Talkback: Report of Public Discussions.** Wellington: Governmental Printing Office.

Pennington, F.C., 1985, Needs Assessment in Adult Education. In Husen T. and Postlethwait, T.N. (eds.) **International Encyclopedia of Education.**

Postlethwaite, T.N., 1985, Error Analysis in Curriculum Evaluation. In T. Husen and T.N. Postlethwaite (eds.) **International Encyclopedia of Education.**

Rieken, H.W., Boruch, R.L. et. al., 1974, **Social Experimentation**. New York: Academic Press.

Rosenblum, Y., 1973, Comparing Discriptive Scientific vs Narrative Style in Science Textbooks. In Lewy A. (ed.) **Studies in Curriculum Evaluation.** Jerusalem, Ministry of Education, Curriculum Centre.

Rosenshine, B., 1985, Direct Teaching. In Husen T. and Postlethwaite T.N. (eds.) **International Encyclopedia of Education.**

Sanders, J.R. and Cunningham, D.J., 1973, A Structure for Formative Evaluation in Product Development. **Review of Educational Research**, 43, 217–36.

Scriven, M., 1967, The Methodoloy of Evaluation, in Tyler R.W., Gagne, R.M. and Scriven, M. (eds.) **Perspectives in Curriculum Evaluation**. Chicago: Rand McNally.

Stake, R., 1967, The Countenance of Educational Evaluation. **Teachers College Record,** 68, 523–540.

Stake, R., 1977, **Formative and Summative Evaluation** In D. Hamilton et al (eds.) **Beyond the Numbers Game** London, Macmillan Educational

Stufflebeam, D.L. et al., 1971, **Educational Evaluation and Decision Making**. Itasca, IL: Peacock Publ.

Suarez, T., 1981, Needs Assessment. In Lewy A. and Nevo D. (eds.) **Evaluation Roles in Education**. New York: Gordon and Breach.

Tamir, P., 1981, Evaluating Curriculum Implementation. In Lewy A. and Nevo D. (eds.) **Evaluation Roles in Education.** New York: Gordon and Breach.

Walberg, H., 1974, **Evaluating Educational Performance**. Berkely, McCutchan.

Walker, D.F. and Schafferzik T., 1974, Comparing Curricula. **Review of Educational Research**. 44, 83–111.

White, R.T. and Gagne, R.M., 1978, Formative Evaluation Applied to a Learning Hierarchy. **Contemporary Educational Psychology** 3, 87–94.

White, R.T., 1974, The Validation of a Learning Hierarchy. **American Educational Research Journal**, 11, 121–136.

Willis, G. (ed.), 1978, **Qualitative Evaluation**. Berkeley: McCutchan.

Zuckerman, J.V., 1954, Predicting Firm Learning by Pre-release Testing. **AV Communication Review**, v.2, 49–56.

Chapter Two

Evaluating Curriculum Implementation

The 1950s and the 1960s are frequently referred to as the decades of educational reforms. Large-scale innovative curriculum projects, fostering programs and organizational changes were introduced to several educational systems throughout the world (Kliebard 1985). Research and development agencies invested time and energy to determine the desired parameters of the proposed changes and to produce materials, guidelines, blueprints, etc. which, according to their view, were needed to ensure the successful implementation of the planned reforms. In the 1970s questions arose about the success of these reforms and their contribution to the attainment of desired goals. The focus of interest switched from specifying the parameters of the desired change to examining their fulfillment in actual educational practice. McLaughlin and Marsh (1978) summarize their views by writing that an important lesson of the so-called decade of reform is that even the best educational practice is unlikely to fulfill its promise in the hands of an inadequately trained or unmotivated teacher. But the failure to put into practice innovative educational programs should not be attributed only to the lack of adequate staff preparation. Experts of communication theory and system theory call attention to a series of factors which may hinder the diffusion of innovation: satisfaction with the existing situation, the lack of clarity with regard to the goals of the innovation, poor communication between the change agent and the user, etc. The actual use of certain innovative educational programs or curricula is usually labelled "curriculum implementation". A more formal definition is: "putting into practice the essenial features of an innovation" In 1977, the following definition, was offered: "implementation is the actual use of an innovation or what an innovation consists of in practice" (Fullen and Pomfret, 1977). This definition is not merely a stylistic refinement of the previous ones, but it is conceptually new. It broadens the scope of the concept "implementation". Here, implementation does not perforce mean following all the prescribed guidelines of an innovative plan. In other words, it does not require absolute fidelity to the original intent of the developers of the innovation, but denotes the adaptation of the innovation to the reality of the absorbing system.

The definition of Fullen is compatible with the contention of Stiles and Robinson (1973) that implementation is the process of enacting changes that others have instigated. They claim that it is a myth that educational administrators, supervisors, teachers and professors are leaders of change, despite the fact that they are always involved when changes occur. Normally they implement or carry out changes imposed by outside forces. The chances of failure are so great that most educational professionals choose to react and implement, rather than to lead change.

A distinction between the innovative program itself and the process of its implementation is emphasized by Goodlad and Klein (1970), who called attention to the eventual gaps between the intended program and its realization in the class. Fullen and Pomfret (1977) list four reasons why implementation should be studied. First, to know better what indeed has changed, and to avoid viewing the innovation as a "black box" without specific information about its actual parameters. Secondly, one must understand the reasons why so many educational changes fail to become established. It is essential to find out which features of the information can, and which cannot, be realized in practice. Third, without providing evidence that a program (or treatment) was adequately implemented it is useless to examine its effects, since failure to attain program goals may be due to the lack of adequate implementation. As Charters and Jones (1973) wrote:

> "If the difference between what the research regards as "experimental" and "control" programs is fictional, the findings of the study will be fundamentally misleading irrespective of their nature. This logic, though simple, is often overlooked both in conducting and interpreting evaluative studies". (p.5)

Fourth, it is necessary to examine whether the differences in outcomes across schools are related to differences in the degree of program implemntation. The arguments put forward by Fullen and Promfret (1977) were reinforced by numerous implementation studies in the 1980s. Wang et al (1984) indicate two reasons for studying program implementation: First, to describe program effects adequately, data are required on how the program was implemented. Second, successful implementation requires explicit information about the program's operating features and implementation conditions. Also, Scheirer and

Removic (1983) argue that...

> "the concept of 'degree of implementation' is critical in order to derive valid conclusions from both outcome and process studies of social and technological innovations. To correctly attribute the observed outcomes of a social program to the intervention, the reseacher should have empirical evidence on the extent to which program components were implemented. Without such evidence, researchers may erroneously conclude that an intervention was ineffective when, in fact, treatment implementation was inadequate to afford a valid test of the program. Further, even successful interventions may not be replicable without a knowledge of program components (p. 599)".

In spite of the recognized importance of implementation studies, educational evaluators have not developed a comprehensive conceptual framework for carrying out such studies, nor have they specified standard methodological paradigms for constructing measurements of implementation. The conceptual framework of implementation evaluation is rooted in the following areas of scientific inquiry: curriculum theory, system theory, innovation studies, diffusion theory and evaluation theory. Terms borrowed from all these domains are frequently encountered in implementaion evaluation studies. Schematically, one can define the location of implementaion evaluation studies as the intersection of these domains. Figure 2.1 is an attempt to graphically represent the relationship of implementation studies to these domains of inquiry.

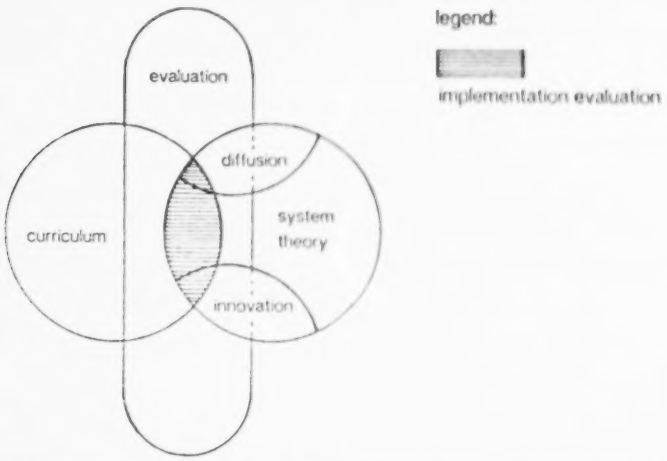


Figure 2.1 The field of evaluating curriculum implementation.

Innovation studies and diffusion theory are presented in Figure 2.1 as subsections of system theory. In each of the domains appearing in fig. 2.1, one encounters a variety of structures, conceptual frameworks, approaches, and so on. The planner of the implementation evaluation study must therefore first select a particular set of concepts from each of these domains and then establish an integrative framework for linking them together in the context of a unique situation.

In order to describe the present state of implementation evaluation studies, a classification scheme in the form of a mapping sentence will first be offered. Explanations will then be offered about the relevance of the concepts appearing in the mapping sentence to implementation evaluation. Finally, some remarks about the nature of such evaluation studies will be presented.

A Classification Scheme for Evaluation Studies Dealing with Curriculum Implementation

The classification scheme presented here is formulated in the shape of a mapping sentence developed by Guttman (1970) as a notational device for research design. One objective of a mapping sentence is to analyze a complex concept and to describe its component elements.

Fig. 2.2 contains a mapping sentence definition of implementation evaluation. As illustrated in figure 2.2 a mapping sentence contains several facets and each one describes a particular dimension of concept. Thus, for example "program orientation" and "change scale" are dimensions of implementation evaluation. The mapping sentence describes a great variety of implementation evaluation patterns. Reading the mapping sentence one always selects only one relevant element from each facet.

A sentence of this type represents one particular evaluation pattern.

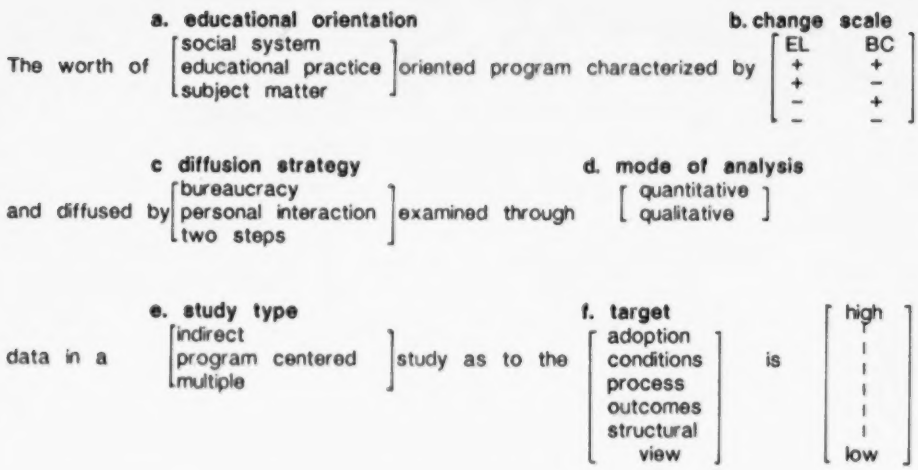


Figure 2.2 Mapping sentence of curriculum implementation evaluation

To illustrate the meaning of such sentences let us characterize an instance of an implementation evaluation study. Adar and Fox (1978) evaluated the implementation of a history curriculum in the Israeli school system. Let us examine it according to the mapping sentence. The curriculum was based on the principle of inquiry retraining and its implementation required the retraining of teachers and coordinating activities with school principals and supervisors. Coordination with the Educational Television Network programs was also needed. The diffusion of the program was done through a central govenmental agency which has the authority to make decisions about school programs in the whole system. The evaluators collected data through observing teaching-learning activities, and examining to what extent teachers analyze historical sources appearing in the textbook. They calculated the percentage of teachers who devote time to both analyzing sources, and to discussing the credibility of those sources.

Utilizing the categories of the mapping sentence, one may describe these studies in the following sentence:
In a **subject-matter-oriented** program, characterized by demand for **great behavioral** change and by **a great number of change elements,** which was diffused through a **bureucratic structure,** the **classroom processes,** as summarized in a **focused study** through a **quantitative data summary mode,** were found to be not fully satisfactory.

Descriptive Dimensions

This section contains a description of the facets appearing in the mapping sentence and of their elements. Examples of empirical studies will clarify the relevance of these facets to differences among evaluation approaches.

Two facets of the mapping sentence describe characteristics of the innovation being evaluated: The educational orientation, and the scale of the change. The third facet, i.e. diffusion channel, has its roots in communication theory. Three other dimensions of evaluation studies may serve as an appropriate framework for characterizing approaches to implementation evaluation: the mode of the data summary; the type of the study; and the target of the study.

Educational Orientation

One may observe a great variety of innovative curriculum projects. They differ with regard to the nature of learning objectives, content, learning strategies, structure, and the type of media used. To describe problems encountered in implementation evaluation it seems useful to distinguish between three types of curricula, based on the nature of the demands leading to their development.

On the basis of Hampson's (1973) theories one can distinguish between innovative curricula of three different eduactional orientations: Social System Orientation (SSO), Practice Orientation (EPO) and Subject Matter Orienation (SMO).

Social System Oriented (SSO) Programs – Some innovative educational programs focus on serving social goals, such as increasing the competency level of skilled workers in industry, motivating the learner to study technological subjects and creating better understanding between various social groups. They may emphasize the enhancement of certain values, such as strengthening feelings of national unity, encouraging participation in community affairs, improving the quality of life, etc.

Innovative programs of the SSO type may demand far-reaching changes in the whole educational program, and are likely to succeed only if extra-school factors provide great support to the school in attaining these goals. Evaluation studies of these innovations ought to concern themselves with issues such as the support of various educational agencies (community, church, mass communication media, voluntary social organizations, etc.) and with the long-range outcomes of the school program.

Educational Practice Orientation (EPO) Programs: Some innovative programs have an educational practice orientation. Their goal may be: the individualization of learning, the teaching of small groups or heterogeneous classes, or the utilization of technological teaching aids such as programed instruction or computers, and so on. These are examples of organizational changes in the learning environment. But EPO programs may also concern themselves with intellectual processes such as learning by inquiry, independent study, etc.

The implementation of such programs requires the coordination of activities primarily within the school site, and is less dependent on extra-school factors than programs pertaining to the previous category. The significant outcomes of such programs are mostly long-range changes in the learners' behavior and some proximal goals such as satisfaction, participation and involvement in activities compatible with program goals.

Subject Matter Orientation (SMO) Programs: Programs may focus on changes in the teaching of certain subject-matter in the school. Several reasons may justify the need for producing a new curriculum of this type. Thus, scientific development in the domain of the subject may make existing textbooks obsolete, writers may wish to employ more adequate methods of presenting materials than appear in available textbooks, or new subject matter may be intruduced into schools for which books are not yet available, such as astronomy, geology, mass media, film studies, etc.

The sucess of programs or this type may be dependent on the within-school and between-school cooperation of teachers teaching a certain subject or on adequate staff training or retraining, and the significant outcomes of such programs may be the mastery of specific information and of well-defined intellectual skills.

The Scale of Change

Havelock (1977) introduced the idea of "the change scale". He writes:

> Schematically when we refer to differences in scale in projects we mean two things. First, we mean a project which has a greater or lesser number of elements in it: people (who teach, who are taught, administrators,...experts, etc.), materials and other technical resources, and what we have called connexions (plans, communication, decisions and coordinating arrangements) which connect the people with one another and with the materials. Secondly, these may be projects...which require people to behave in ways which are more or less radically different from their customary behavior...
>
> (p. 85)

This distinction yields the following permutations:

(1) Large scale in elements (EL+), large scale in changing people's behavior or attitudes (BC+).
(2) Large scale in elements (EL+), small scale in changes (BC–).
(3) Small scale in elements (EL–), large scale in changes (BC+).
(4) Small scale in elements (EL–), small scale in changes (BC–).

The "educational orientation" of a program, as described in the previous facet, is orthogonal to the "scale of innovation" facet. This means that in practice one can imagine all possible combinations of the elements of these two facets. Educational programs may demand changes in the following areas: The content taught in school, the teacher's role, the hierarchical structure of the staff, the teacher-learner relationship, etc. They also may require coordination between various elements, such as school and extra-school activities. Thus, for example, a decision to increase the extent of contemporary history or local history within the framework of the history curriculum may constitute a "small scale in elements (EL–) and small scale in behaviour changes (BC–)" type of innovation, if adequate curricular materials are readily available within the system. On the other hand, the transition to a problem-oriented intergrated Social Studies Curriculum, from a strictly discipline-oriented educational program in which History, Geography, Economics, Civics, etc. are separately taught by subject matter experts, may be both large scale in elements (EL+) (retraining teachers, producing new materials, changing the examination system, creating consensus, etc.), and large scale in behaior changes (BC+).

The Diffusion Strategy

The diffusion pattern of innovation is another facet which characterizes program implementation and may affect the nature of the evaluation study. In an educational system where decisions about educational programs are made by central authorities, there is less need to study the adoption pattern of the new program than in a system in which each school is allowed to make desicions about curricula. Rogers and Shoemaker (1971) mention three traits of any innovation, which are related to diffusion: complexity, triability (in an earlier version the term "divisibility" appears instead of triability), and observability. As to the diffusion channels, a distinction is made between interpersonal and mass-media channels. Observing diffusion practice, we distinguish three patters: bureaucratic diffusion, personal interaction diffusion, and a combination of a centrally instigated mass-media campaign with a personal contact approach.

Bureaucratic Diffusion Pattern – One form of program diffusion is the bureaucratic pattern, which usually operates through hierarchically related diffusion centers. A main office coordinates several regional centers and the regional centers delegate reponsibility into the hands of district centers, etc. The major means of dissemination are messages produced by the main office which are transmitted to the users directly or through the regional or district offices. The channels of communication may be printed materials (teacher's guide, newsletters, circulars), broadcast transmissions or visiting lecturers sent to regional teachers' conferences. Bureaucratic diffusion patterns need not necessarily operate through non-personal communication, nor need they be engaged in a unilateral communication with the program users. Frequently, diffusion offices employ change agents whose task it is to contact users and to persuade them to adopt the innovation. These agents may also collect feedback information from teachers directly or through questionnaires. Nevertheless, the limited amount of personal communication and the utilization of routine forms (questionnaires, etc.) severely restricts the transmission of complex messages.

Personal Interaction Patterns – Some curricula are designed in such a way that their implementation presupposes personal contact with the teachers, which may develop into some form of social interaction. Frequently such contact is already created at the stage of program development. Teachers themselves may or may not participate in writing, trying out and revising the program, but at the stage of implementation "personal interaction" with teachers becomes an absolute necessity. In such a setting, the development team, together with specially trained mediators or guides discuss problems of utilizng the program with classroom teachers. They try to locate misconceptions and errors occurring during program implementation and to suggest ways for overcoming difficulties.

Unlike the bureaucratic diffusion pattern in which feedback from teachers is obtained through standard data collection devices such as questionnaires or "mass conferences", the Personal Interaction Pattern data are collected through interviews, small group discussions, the analysis of actual video-taped lessons, etc. These procedures enable the transmission of more complex ideas to the teachers and are apt to bring about changes in their behavior. Accordingly, curriculum writers using this pattern of contact may feel themselves free to include types of activities in the program which may require the substantial reorientation of teachers.

The Two-Step Flow Model – Communication studies found that the mass media by themselves are not powerful enough to diffuse innovations. Only the combination of mass media messages with person-to-person communication among peers brings about the widespread diffusion of innovations.

The information about an innovation first reaches the "opinion leaders", i.e. person whose opinion is highly valued by people, and through personal communication these may then motivate people to adopt the innovation. Evidence about the effectiveness of two-step communication flow model is summarized by Katz (1957).

The Mode of Data Analysis

Evaluation methodology has been greatly influenced by the tradition of quantitative data summary. The majority of educational data has been collected in a way which yielded only a quantitative index for some abstract trait. Traits such as "reading comprehension" or "school attitude", etc., are described by some numerical score. Measurement instruments have been developed to yield a quantitative expression of certain behavior traits. Even the direct observation of the teaching process or the critical judgement of a certain text has been transformed into a quantitative result by the means of observational models or content analysis schemes. Recently, great interest has been revealed in qualitative data summary. The observational methodology successfully used in ethnographic studies found its way to evaluation studies in education. Stake (1978) claims that the prose description of events offers explanatory power beyond that which is available from using statistical procedures alone. Guba (1978) outlines a methodology of Naturalistic Inquiry, and explains the advantages of this methodology within the framework of educational evaluation. A claim for using qualitative date summary in evaluation studies has been made by several researchers (Lewy 1977) and an excellent example of such a study is presented by Smith (1978).

The Type of the Study

One may observe a wide variety of implementation evaluation study types. Three types will be described here: indirect study, program centered study and multiple studies.

Indirect Studies – One approach to evaluate the implementation of innovative curricula is to conduct a comprehensive study about the state of the art in a specific curricular area. Studies of this type are not designed to deal with the implementation of one particular curriculum directly and therefore we refer to them as indirect studies. Information touching on one particular curriculum is a by-product of such studies. Several examples will illustrate the great variety of studies included in this category. Some, for example, represent a large-scale examination of many aspects of an educational system. Thus, the IEA (International Association for the Evaluation of Educational Achievement, located at the University of Stockholm, Sweden) studies of mathematics (Husen 1967)

and of science (Comber and Keeves 1973) pertain to this category. The IEA studies examined educational achievements in twenty-one countries and provided information about the scope of the utilization of innovative programs and the achievement level of those students who studied according to these innovative programs.

Another general survey type of study is the series of Case Studies in Science (National Science Foundation 1978). That study focused on science teaching in the United States, and through examining the situation in the whole system, it made an attempt to provide information about the spread of innovative science programs, and their impact on school practice. Surveys need not necessarily be large-scale studies. An example of a small scale indirect evaluation study of implementation was carried out over a period of six weeks. It examined the spread of innovative curriculum materials used in Israeli schools (Lewy 1979). In this study a brief questionnaire was administered to a sample of school principals. The principals were asked to indicate what curriculum materials were used in their school in the major subject matter areas, and also to express the level of satisfaction with new curricula disseminated in the system. The study provided information about the diffusion of innovative programs and about satisfaction with these programs.

Program Centered Studies – In this category are included all those studies which focus on the implementation of a particular program and on the changes occurring in the school and outside the school as a result of its use. Despite their variety, most implementaion evaluation studies belong to this category.

An example is Sabar's (1986) study. She observed the implementation of a biology program, recorded class-room events, and then identified program components which were well-implemented (in more than 70 percent of classes) and others, which were not well-implemented (in less than 50 percent of classes), (see Table 2.1).

It can be seen that objectives related to general biological principles such as "adaptability of the organism to the environment" or "similar solutions to different problems" were well implemented in the classes. Also, other innovative features of the program, such as "manipulating experiments" and "discussion" were fully implemented, while the topic "environmental equilibrum" had a low level treatment in the classes. The author presented a series of suggestions which has implications for improving the implementation of the particular biology program.

Multiple Studies – In several instances, the implementation of innovation has been examined by a series of studies, each one addressing itself to a different aspect of the program utilization. Atkin (1979) writes about three studies, each one employing a different approach, which were simultaneously conducted to assess the current status of theaching science in American schools. One of these approaches was the previously mentioned case study series, another was a survey research, an the third an analytical review of the literature. He concludes that "thus by triangulating on science teaching, using three different approaches, it was hoped that a reasonably accurate picture might be developed as a benchmark of the current state of science in the schools". (p.397).

Multiple studies may be simultaneously planned and carried out, or alternatively, they may be conducted without preliminary coordination. Thus it happened that dozens of evaluation studies were conducted on the Follow Through Program (Anderson, 1977; House et. al., 1978). Such studies may be summarized through various techiniques of meta-analysis (McGaw, 1985).

The Targets of the Evaluation Study

Implementation is a complex process and the evaluation may decide to focus his attention on some particular phase or component of the process, or, alternatively, to deal with the whole process and to examine the inter-relationship between its component elements and with program outcomes.

We will present examples of studies related to various evaluation targets.

Table 2.1 Percentage of teachers observed implementing subvariables

Variable	Sub-variable
	High Frequency
Objective	Adaptability of organisms to the environment (81%)
	Similar problems & different solutions (81%)
	Acquisition of biological concepts (81%)
	Planning of data gathering (90%)
Materials	Text books (97%)
	Work sheets (97%)
	Living organisms (86%)
Activities	Discussion (100%)
	Manipulating eperiments (91%)
	Summing experiments (75%)
Stategies	Group work with same tasks (75%)
	Self study (78%)
	Teacher's demonstrations (72%)
	Low Frequency
Objectives	
Ideas & principles	Environmental equilibrium (25%)
Inquiry	Understanding scientific processes (34%)
	Distinction between facts (40%)
	Defining problems (50%)
Affective	Care & respect for animals (37.5%)
Materials	Auxiliary books (47%)
Activities	Field trips (25%)

(Sabar 1986)

Adoption of Innovations – Several studies examine the diffusion or the pattern of adoption of innovations. Lewy (1979) and Owen (1979) describe the diffusion of innovative programs in Israel and Australia, respectively. The persistence of adoption is another important issue in evaluating implementation. Welch and Ward (1975) defined persistence as using the curriculum a second year, and found that persistence of adoptation across innovative curricula varied from 37% to 82%, with an average of 64%.

A refinement of adoption studies is observed in the attempt to establish a measure for the level of program utilization. This attempt constitutes a shift from the dichotomous distinction of adoption and non-adoption, and enables one to classify adopters on a quantitatively structured scale. Hall and Loucks (1977) developed a scale called Level of Use (LoU).

Table 2.2: Levels of Use of Innovation (LoU)

Level of Use	Definition of Use
0 Nonuse	The user has little or no knowledge of the innovation
I Orientation	Takes action to learn more detailed information
II Preparation	User is preparing first use
III Mechanical	Focus effect on short term, day-to-day use
IVA Routine	Use is stabilized
IVB Refinement	User varies the use
V Integration	Combines own efforts with related efforts of colleagues
VI Renewal	Reevaluates, seeks modifications or alternatives

Implementation Conditions – Numerous evaluation studies concern themselves with the conditions of implementation. Fullen and Pomfret (1977) refer to these conditions as determinants of implementation and warn against confusing determinants and the actual process of utilizing the innovation.

The theoretical basis of these studies is derived from communication theory, system theory, organizational management, staff development, and so on. Havelock (1977) describes innovation processes in terms of system theory, and offers three basic concepts necessary to investigate and to explain implementation conditions: **Infrastructure,** such as roads, the transportation system, the communication system, knowledge dissemination institutes, the capacity to

use resources, and the awareness of needs; **Authority,** including leadership, connections and cohesions in the hierarchical structure of the bureaucracy; and **Consensus**, related to the value of the innovation. Utilizing these key concepts he developed a questionnaire in which he examined the existence of potential difficulties in the course of carrying out a project. In this questionnaire, the questions are clustered in groups, according to specific factors: geographic, historical, social and cultural, political economic, procedural and personal. Some illustrative items included in the questionnaire are presented in Table 2.3.

Table 2.3: Potential difficulties

	Extremely serious	Very serious	Somewhat serious	Observed but not serious	Not observed by me
6.14 Ruling elites traditionally opposed to innovation or to expansion in education sphere					
6.15 Struggle for independence left the educational structure in disorder					
6.16 In the struggle for independence many technically qualified teachers and teacher educators left the country					
6.17 Other historic factors (specify and rate seriousness)					
6.20 Social and cultural factors in general					
6.21 There was conflict among ideologies about change					
6.22 There was a poor climate for sharing ideas openly					
6.23 There was inadequate social harmony and good relation among project team members					
6.30 Political factors in general					
6.31 Disputes among diferent political parties or factions					
6.32 Objections to project by special interest groups					
6.33 Lack of interest by political leaders					
6.34 Cartain key political leaders were opposed to project					
6.50 Economic factors in general					
6.51 Indequacy of level of financial aid from outside the country					
6.52 Indequacy of level of financial sopport contribued by the country					
6.53 Inflation threw original cost estimate completely off					
6.54 Costs were generally underestimate even without considering the effect of inflation					

A second focus of implementation conditions is the issue of staff development. The Rand Corporation study (McLaughlin and March 1978) addresses itself to this question. This study identified clusters of variables which are considered to be essential in facilitating the successful implementation of new programs:
(1) Institutional motivation – This can be enhanced by collaborative planning activities, i.e. by the involvement of teachers in the planning activities from the very beginning of the development, by creating intrinsic motivation for executing the changes, and by active involvement of the district managers in the dissemination activities.
(2) Project Implementation Strategies – These include (a) staff training activities which may be either skill-specific, or more concerned with the teachers decision-making ability; and (b) staff support through consultants, peer discussions and the freedom to modify elements of the program.
(3) Institutional Leadership – The support of the institution's principal and a good working relationship among teachers within the institute are both significant contributors to good program implementation.

User's Satisfaction – Several different groups can be considered as the users of a curriculum: the learners, the teachers, the school principals or the supervisors, etc. The implementation evaluation may concern itself with the satisfaction of any or all of these user groups. In this section, we present examples of studies done to the teachers' satisfaction.

Welch (1975) developed a Likert-type questionnaire called the Curriculum Attitude Survey (CA) to measure various components of teachers' attitudes. The items of the questionnaire are related to general features of satisfaction such as "cooperation with other users", "self confidence in using the program", "identification with other users", etc. The questionnaire is not program-specific, i.e. it can be used to measure satisfaction with a variety of innovative programs. A different approach is employed in Fraser's (1978) study. He posed program-specific questions to measure mathematics teachers' attitudes toward a mathematics game. The teachers were asked to express their opinions about the unique features of that particular program, in such areas as: the "humanizing" of mathematics: the demonstration that mathematics has practical implications; the arousal of interest in mathematical topics and so on.

Processes – Fullen and Pomfret (1977) consider that a detailed examination of what goes on in the classroom in an innovative program is, by definition, the evaluation of implementation. The comparison of actual events with planned or prescribed ones, is labelled as the fidelity criterion of implementation. A more open approach focusses on "what an innovation consists of in practice" or on the totality of events occurring or not occurring as a result of the adoption of an innovation. Fullen and Pomfret (1977) label this approach to implementaion evaluation as the **process-criterion** or the **adaptive-criterion** approach. Two arguments may justify the adaptive criterion approach to implementation evaluation. Firstly, evaluation should be concerned with unintended and unplanned occurrences as well as with planned ones. Some innovations may have also dysfunctional effects which cannot be easily predicted. This does not mean that side effects or unintended outcomes are necessarily dysfunctional. Indeed, one can point to several beneficial uninitended outcomes of innovation. The need to assess unplanned outcomes of curriculum changes is well-emphasized in curriculum evaluation literarure (Scriven 1972). A second reason for employing an adaptive-criterion approach is that good implementation requires adapting both the materials and classroom behaviours to the needs of the teaching-learning situation. Accordingly, the success of implementation is not ensured by following prescribed instructions exactly, but rather by adaptation to situational needs. Thus the evaluator has to record what the "innovations consist of in practice", and on the basis of this record it becomes possible to judge the merits of the innovation.

Some examples of structured classroom observations to evaluate program implementation will be presented here. Both overall rating of programs and systematic records of events will be illustrated.

Overall Rating: One approach to evaluating classroom processes is to observe the events and to make an overall judgment about the degree of implemetation. Thus, for example, Lukas and Wohleb (1973) report that in the evaluation project of the Head Start Planned Variation program, the sponsors of the program observed classroom procedures and rated the implementation level on a scale from zero to nine (0–nonacceptable; 9–outstanding).

The Systematic Record of Events: Another approach to evaluating classroom procedures is the keeping of records about events according to some previously structured scheme. The recording may be carried out through observation of events or, alternately, through the self-reporting of teachers or learners. It is impossible to summarize the great variety of observation schemes, and no attempt will be made to even present a representative sample of such schemes.

A common characteristic of these methods is the attempt to produce a quantified summary of the records. The quantification focuses on those aspects of the teaching-learning process which are considered to be crucial to the innovative program, or which are the concern of the dominant educational philosophy.

(1) Classroom management – Several new curricula are concerned with changing the classroom setting or the organizational structure of teaching. Individualized instruction, small group teaching and teaching in heterogeneous classes are a few examples of innovative program goals. The utilization of such observational categories is relevant for measuring the implementation of Educational Practice Oriented innovative programs. An example of an observational scheme which focuses on management issues was developed by Levin (1979), with the aim of measuring the implementation of an individualized science program in elementary schools. During a single study period, six learners were selected to be observed. The time unit of one observation was 30 seconds. An observation cycle of all six learners lasted about 3.0 – 3.5 minutes, and during a single period approximately ten such cycles of observation were completed. Three major behavioral categories appeared in the scheme: (a) pupil-teacher interaction, (b) pupil-pupil interaction, (c) pupil-material interactrion. A data summary of Levin's study is presented in Table 2.4.

Table 2.4. Average time engaged in a specific interaction (in minutes)

Level of Achievement	The type of interaction		
	Teacher-learner	Learner-learner	Learner-material
High	1.69	1.96	5.34
Middle	1.69	1.45	6.11
Low	1.46	0.69	6.84

Levin 1979

Since the aim of the innovation was to maximize learner-materials interaction in class, the evaluator perceived the results of the observation as evidence of the success of the program.

(2) Participation – Participation in learning activities, the time spent on tasks, initiating activity, etc., are a few examples of process variables which are of interest for implementation evaluation. Bloom (1954) suggested the method of "stimulated recall" to determine the amount of relevant thinking, as well as to study the kinds of thinking students do with a set of currriculum materials. He claimed that recording events and replaying them to students after the lesson, along with an extensive questioning about what they recalled of their thoughts, enables one to obtain an accurate indicator of student participation in class activities. Later, more direct observations on the student's academic learning time were carried out (Denham and Lieberman 1980). Studies of this type enabled researchers to determine which episodes contained in the lesson generated participation, and which did not.

(3) Intellectual Processes – A third type of process which frequently appears in observation schemes describes the nature of the intellectual activity which takes place during the learning period. In Educational Practice Oriented programs the innovation theme might be "inquiry learning", "problem solving", or "independent work", etc. An example for examining intellectual processes within the framework of a subject-matter oriented innovative program is provided by Adar and Fox (1978). In an innovative history curriculum, for example, the utilization of historical sources as a means of learning about the past is emphasized. The authors of the textbook have not specified what are the parameters of "the use of historical sources". The evaluation team, however, before starting the work of data collection, specified a variety of behaviors which they judged to constitute an adequate approach to handling sources. Among these

specifications were: examining the credibility of the source, providing information about the nature of the source, about its author, time, etc. The specification of these behaviors was considered to be an important step in the evaluation process.

Having specified the parameters of using historical sources, the evaluators observed 72 teachers in their classes using the new curriculum. Ten of them disregarded the sources entirely, while more than half of the teachers treated the sources as a regular reading passage included in the textbook. The latter teachers explained the contents of the source, but did not utilize it as evidence about past events. Only 20 percent of the teachers presented background information about the sources.

Outcomes – In numerous studies, the attainment of the program goals is identified with the notion of program inplementation. This conception of implementation evaluation may well serve Educational Practice Oriented programs. Thus, for example, in a program of individualized learning, one may define the program goals as maintaining individualized learning practice in the class, and consequently the implementation of the program and the attainment of the program goals is practically identical. There are, however, some flaws in this conception. The goals of an individualized program are not the process of individualization itself. Individualization is only a means to attain some agreed learning objectives. Defining the goals of individualized learning in terms of the process of individualization implies that either one considers this to be worthwhile practice – without regard to its effect on the learner's behavior – or else implies that one has a strong conviction and satisfactory evidence about the efficiency of individualization in producing certain desired learner behavior.

In the latter case, the individualization is only a sign of goal attainment, which should not be mistaken for the program goals themselves.

A Structural View of Variables Determining the Implementation Process – In many cases, evaluators made an attempt to include in their studies all those variables which they considered to pertain to the domain of implementation, rather than restrict their interest to some unique features of this complex process. Three studies in which such attempts have been made will be presented here as examples. Goodlad and Klein (1970) referred to implemen-

tation as a series of transitions from a given stage (in the program development and utilization) to subsequent stages. They defined five such stages: (1) the ideal program; (2) the actual production of learning materials; (3) the teachers' perception of the program; (4) the teachers' performance; and (5) the learners' expreience.

According to this approach, the evaluation of program implementation would mean the examination of all possible transitions from a certain stage to another. The Adar and Fox (1978) study followed this approach, though it did not address itself to all possible transitions. Adar and Fox performed a series of congruence studies examining the transition from the definition of objective for a history curriculum, to the actual set of learning materials. Their study focused on determining to what extent the textbooks reflect the principles put forward in the program rationale. They then examined the teachers' perceptions of the materials and the actual classroom behaviors. Interestingly enough, a higher congruence was found between the teachers' actual classroom behavior and the program rationale, than between the textbook and the program rationale developed by the same team. Thus, rather than observing a gradually increasing discrepancy between the ideal program goals and their actualization in the schools, the authors found that during the process of implementing a progarm one may be able to correct flaws in previous transitions.

A time sequence of occurrences constitutes the backbone of another evaluation study in which the author selected a series of major variables, which, according to his conception represented a chain of causation (Silberstein 1980). The first of these variables is the teacher's familiarity with the program, the second, his approval of the progaram rationale, the third, his adherence to the program guidelines in actual teaching situations, and the last would be the cognitive and affective outcomes. A schematic representation of the interrelationships between the variables is presented in figure 2.3

Silberstein tried to control variation due to entry behavior differences between learners by using a test of scholastic achievement. He then applied the path analysis technique in an attempt to explain the residual variation in program outcomes.

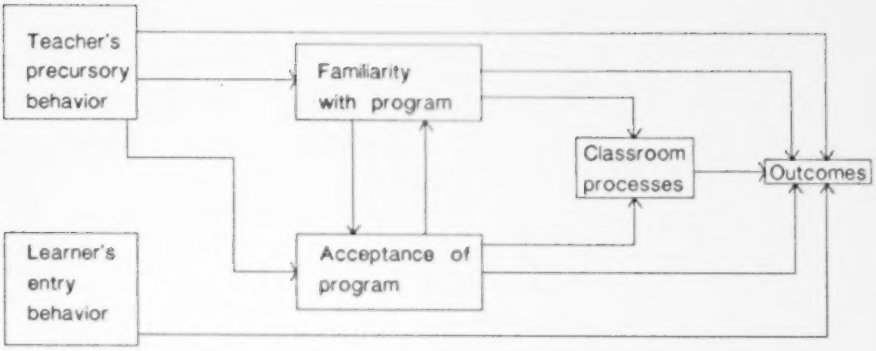


Figure 2.3: A structural model of variables of implementation

The Summative-Formative dimension of Implementation Evaluation

Evaluators tend to associate the differences between formative and summative evaluation with differences between the small scale try-out of innovations and large scale program implementation. It is frequently said that evaluation at the small scale try-out stage should be formative, while that carried out on a large scale is necessarily summative. Such a distinction between formative and summative evaluation is rooted in the misinterpretation of the terms. As suggested by Scriven (1967), the two types of evaluation differ from each other with regard to the roles they are supposed to fill. Thus rather than speak about formative and summative evaluation, it may be more appropriate to speak of the formative to summative decisions taken on the basis of summarizing evaluation data. Formative decisions usually aim to correct flaws and eliminate weak components of an entity. They may touch upon: elements of the instructional materials; arrangements made in a system to disseminate the program; guidelines given to teachers about how to use the program; the content or the schedule of inservice training programs; adjustments made in schools to use the new program; and so on.

In contrast, summative decisions are formulated in terms of "adoption or rejection", and "continuance or discontinuance" of a particular program. Summative decisions touch upon a whole entity or process, rather than upon its com-

ponent elements. Accordingly, in the context of formative evaluation the decision-maker needs detailed information about all components of the evaluation target, while in the context of summative evaluation, the decision-maker faces the problem of weighting and combining evaluative information related to various program elements. Moreover, summative decisions can also be based on a holistic assessment of a particular entity.

An example of utilizing formative evaluation for improving program implementation is presented by Leithwood and Montgomery (1978). To unravel the complexity of the implementation process they pointed out that not only was implementation multilevel, but it was multidimensional as well (Fullen 1981). Adapting a sociological model to curriculum innovation, they identified eight implementation components: global conception; objectives; student entry behaviors; content; instructional materials; teaching strategies; learning experience of students; assessment tools and procedures.

The next step was to measure the Level of Use by a particular teacher, or a group of teachers, of each of the eight curriculum components. In this way they were able to plot "implementation profiles" of teachers at the individual, subgroup, and group levels.

For example, figure 2.4 reports some results on five selected dimensions for a group of 30 primary school teachers using a science curriculum.

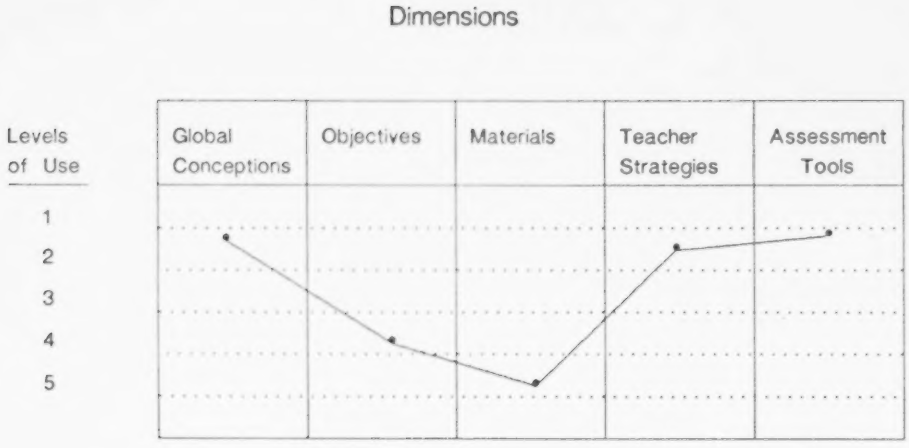


FIGURE 2.4: Innovation profile of 30 science teachers according to Levels of Use on five curricular dimensions

Such profiles can be used as input for formative evaluation decisions concerning program implementation. Decisions can be made to focus in-service training activities on components which have not been sufficiently used, or to adjust the system in a way which may increase the utilization level of these neglected components.

Profiles may also serve as input for summative evaluation. In this context one has to decide whether given forms of profile variation are acceptable, and what kinds of deviations from the ideal implementation pattern may be permissible.

Measuring Degree of Program Implementation

Methods used to measure program implementation are reviewed by Scheirer and Remzovic (1983). The authors classify implementation measurement tools into categories which are roughly ordered from the more objective (not requiring human judgement) to the more subjective (dependent on human observer or respondent). A slightly abbreviated list of categories suggested by the the authors appears in Table 2.5.

Table 2.5: Techniques for measuring

Extent of Human Judgment Required	Technique
low ↑	Equipment performance
	Behavioral observation
	Institutional records
	Interviews
	Questionnaire
↓ high	Ethnographic observation

(Scheirer and Rezmovitz 1983)

A brief description of these techniques is presented below:

Equipment performance – This category of measures refer to direct observation of whether a particular equipment operates correctly. In evaluating a computerized course, a course based on videotape records, or a course which requires the utilization of a microscope, one may employ technical devices for measuring implementation.

Behavioral observation – Behavioral observations were defined by the authors as observer-collected data using a prespecified set of categories, each with an operational definition of the target behaviors to be recorded.

An example of such an observational scheme is provided by Leinhardt (1978) in which some of the following issues were observed:

- The number of adults present in the class
- Children decide what units of mathematics to study
- Children get own prescription ticket

Most observational data of this type are collected in explicity stated time units. For example, Yoloye (1977) developed a scheme in which the observer records minute by minute which of the following events occured: teacher's lecture, group activity, or whole class discussion.

Institutional records – Researchers prefer collecting data rather than using archival data. Nevertheless, available institutional data, too, may provide useful sources for evaluating program implementation. Morris and Fitz-Gibbon (1978) provide a list of records often available about educational programs. Their list is reproduced in Table 2.6.

TABLE 2.6 Records often produced by educational programs

- Attendance and enrollment logs
- Progress charts and chcklists
- Certificates of completion of activies
- Sign-in and sign-out sheets
- Parental permission slips
- Completed student workbooks
- Logs, journals, and diairies kept by students, teachers, or aides
- Unit or end-of-chapter tests
- Teacher-made tests
- In-house memos
- Newspaper articles, news releases
- Dog-eared and worn textbooks
- Flyers announcing meetings
- Circulation files kept on books or other materials
- Records of bookstore or cafeteria purchases or sales
- Legal documents such as licenses insurance policies, rental agreements, leases
- Diplomas or transcripts
- Report cards
- Letters of recommendation
- Bills, purchasing orders and in voices from commercial films providing goods and services
- Activity or field-trip rosters
- Student assignment sheets
- Products produced by students, such as drawings, lab reports, poems, themes and essays
- Letters to and from parents, business persons, the community

They claim that such records enable one to reconstruct a credible portrayal of what had gone on in the program. Archival data are considered to be non-reactive, and therefore they are characterized by a high level of credibility. The utilization of institutional records however, may entail several problems. They may lack accuracy, their significance may be misinterpreted by those who are not familiar with the institutional mores, and access to such data may require special arrangements (Luckey, Broughton and Sorenson 1982).

Interviews and questionnaires – The most common technique to measure program implementation is to ask people who are involved in the implementation process. Problems related to using such information-gathering techniques are thoroughly discussed in books on research methodology, and there is no need to deal with them here.

Ethnographic observations – The least structured way of studying program implementation is the ethnographic observation, which is also called case study method or naturalistic observation. It proceeds without prespecified categories of observation and puts emphasis on understanding the totality of the situation. It may include a variety of additional methods such as document reviews, interviews, etc. Parlett and Hamilton (1977) provide a clear view of unstructured evaluation designs. They claim that the role of evaluation is to illuminate a new program in use, and they suggest employing a multimethod approach based on participant observation, interviewing participants, documentary and background analysis, and questionnaire and test data.

Implementation as an Independent or as a Dependent Variable

Implementation is treated in evaluation studies both as a dependent and as an independent variable. In the first case the researcher may deal with questions of the following types: Which programs are likely to be implemented (and under what conditions)? What are the diffusion-milieu variables which facilitate implementation? In other words, one may look for reasons to explain the success of implementation. Alternatively, one may be interested in the implementation as an independent variable which may explain variations in the program outcomes. In this case the question to be posed is as follows: What com-

ponents of program implementation are related to a certain type of program outcome?

For the practitioner both types of evaluation questions are of great interest. Therefore, in an actual situation of introducing a certain innovation, it may be desirable to perform studies of both types. Nevertheless, it is possible to identify situations in which one of these study types is of greater interest. Thus, for example, in education practice oriented programs one may give preference to studies in which implementation is considered as a dependent variable. With regard to such programs, one seeks evidence about whether the suggested changes have been put into practice. Since the program goals may be defined in terms of processes or transactions occurring in the class, the program user will not be concerned with examining the outcomes separately from the implementation processes.

Thus, for example, for a program designed to increase the learners' freedom of selecting learning assignments, one would like to examine whether such freedom is exercised in the actual class situation. Due to the overlap between the program goals and the classroom procedures, the implementation evaluation can indicate satisfactory evidence of the program success. But there are at least two other cases when the interest is focused on implementation as a criterion (or dependent variable). The first case is when the program goals are remote ones, and are related to behaviors which will take place years after utilizing the program. Thus, for example, goals of a school program may be defined in terms of adult behavior (such as participation in communities, family life, etc.), or social-system oriented goals, such as training manpower for meeting the demand of shifts in the occupational structure in a given system. In those cases, the examination of program outcomes can take place only in the remote future, and therefore the program user has to content himself with evaluating the implementation. It should be noted, however, that for programs having remote goals one may determine intermediate program goals as well, such as the acquisition of certain skills or of certain attitudes which may serve as indicators of the success of the program.

The second case is when sufficient evidence is available to prove that carrying out a series of activities indeed facilitates the attainment of certain desired educational goals. Thus, for example, if one has sufficient evidence that the

provision of diagnostic feedback increases the attainment of certain educational achievements of the cognitive type, one may restrict one's interest to examining whether such feedback has been provided for the learners. The relationship between the implementation of the program and the desired outcomes is taken for granted, and need not be examined anew. Implementation as an independent variable appears in evaluation studies to explain variances in program outcomes.

The Openess of the Evaluation System

The thrust of this paper has been to provide a conceptual framework for dealing with implementation evaluation. The mapping sentence suggested for classifying these studies may serve as basis for seeing similarities and differences between studies carried out in this field. However, the suggested framework may call attention to types of studies still missing at present in evaluation practice. The suggested classification scheme is only one attempt at classifying evaluation studies, but no matter which classification scheme one would employ, the field can be characterized by the existence of a great variety of approaches. This fact supports the contention that evaluation theory is not mature enough to subscribe to a limited amount of models or schemes (Linn 1978). While this seems to be very true, one may nevertheless pose the question as to what should be, or whether there should be any limits for absorbing new theories, models and approaches into the domain of evaluation. In this context it seems appropriate to quote Havelock (1977):

> Openess is likely to come and go in cycles: no system has an infinite capacity to absorb new inputs, and having absorbed one important kind of input, a system may require a "closed period" in which it digests this input and readjusts its internal functioning accordingly.(p.49)

The field of evaluation has encouraged the production of new models, and created a disproportion between the intellectual effort invested in developing models and in using them in practice. Models have been easily discarded to yield their place to a new ones without making a sufficient effort to explore their strengths and limits. It may well be that evaluation had reached such a high saturation level in regard to theoretical input, that without investing more efforts to absorb it and to digest it the functioning of the system will be greatly damaged.

References

Adar, L. and Fox, S., 1978, **An Analysis of the Content and Use of a History Curriculum** (in Hebrew). Jerusalem: Hebrew University, School of Education.

Anderson, Richard B., 1977, The Effectiveness of Follow Through; Evidence fron the National Analysis. **Curriculum Inquiry,** 7, 209–226.

Atkin, Myron, 1979, Case Studies in Science Education, In **Tamir, P. (ed.) Curriculum Implementation.** Jerusalem: Hebrew University, Science Teaching Center.

Bloom, B.S., 1954, Thought Process of Students in Discussion Classes. In French.S (ed.) **Accent on Teaching,** New York: Harper & Brothers.

Charters, W.W. & Jones J., 1973, On The Risk of Appraising Non-Events in Program Evaluation. **Educational Researcher,** 2(11), 5–7.

Comber, L.C. & Keeves, J.P., 1973, **Science Education in Twenty Countries**, Stockholm: Almqvist and Wiksell.

Denham, C. and Lieberman, A. (eds.), 1980, **Time to Learn,** Washington DC: National Institute of Education.

Fraser, Barry, 1978, Use of Content Analysis in Examining Changes in Science Education Aims Over Time, **Science Education,** v. 62, 135–141.

Fullen, M, 1981, Evaluation and Implementation, in Lewy, A. and Nevo, D., (eds.) **Evaluation Roles in Education,** New York: Gordon & Breach.

Fullen, M. and Promfret, A., 1977, Research on Curriculum and Instruction Implementation. **Review of Educational Research,** 47, 335–397.

Goodlad , J. and Klein, F, 1970, **Behind the Classroom Door**, Worthington, Ohio: Jones Publications.

Guba, Egon, 1978, **Naturalistic Inquiry.** Los Angeles: UCLA, Center for Study of Evaluation.

Guttman, L., 1970, Integration of Test Design and Analysis, In **Proceedings of the 1969 Invitational Conference on Testing Problems.** Princeton N.J.: Educational Testing Service.

Hall, Gene E., and Loucks, Susan, F., 1977, A Developmental Model Determining Whether the Treatment is Actually Implemented, **American Educational Research Journal** v. 14, 263–76.

Hampson, David H., 1973, Educational Product Development, Rep. no.11 Raleigh, North Carolina: State University, Center for Occupational Education.

Havelock, R.G. and Haberman, A.M., 1977, **Solving Educational Problems; A Theory and Reality of Innovation in Developing Countries,** Paris: Unesco.

House, Ernest, R. and Gene V. Glass et al., 1978, No Simple Answer; Critique of the Follow Through Evaluation. **Harvard Educational Review,** 48, 128–160.

Husen, T., 1967, **International Study of Achievement in Mathematics; A Comparison of Twelve Countries.** New York: John Wiley & Sons.

Katz, Elihu, 1957, The Two-Step Flow of Communication: An Up-To-Date Report on a Hypothesis, **Public Opinion Quarterly,** v. 21, 61–78.

Leinhardt, Gaea, 1978, Applying Classrom Process Model to Instructional Evaluation, **Curriculum Inquiry,** 8, 155–176.

Kliebard, H., 1985, Curriculum Movements in USA. In Husen T. and Postlethwaite, T.N., (eds.): **International Encyclopaedia of Education.** Oxford: Pergamon Press.

Leithwood, K.A. and Montgomery, D.J., 1978, Evaluating Program Implementation (Unpublished) OISE Ontario.

Levin, T., Liebman, Z. and Amiad R., 1979, Behaviour Pattern of Learners in Individualized Classes (in Hebrew) Tel Aviv University, School of Education, NILI Project.

Lewy, Arieh, 1977, The Nature of Curriculum Evaluation: in Lewy A. (ed) **Handbook of Curriculum Evaluation,** New York: Longmans.

Lewy, A. and Shye, S., 1978, Three Main Approaches to Evaluating Education: Analysis and Comparison by Facet Technique. In Shye, S., (ed) **Theory Construction and Data Analysis in Behavioral Science.** San Francisco: Jossey-Bass.

Lewy, Arieh, 1979, (mimeog.) **A survey of Curricula Used in Israeli Schools** (in Hebrew), Ministry of Education and Culture, Curriculum Center, Jerusalem.

Linn, Robert, 1978, Evaluation of Title I via RMC Models – a Critical View. **Evaluation Comment,** Vol. 5(4), 4–5.

Luckey, J.W., Broughton, A., and Sorenson, J.E., 1982, Archival Data in Program Evaluation, **Evaluation and Program Planning,** V.5 319–26.

Lukas, C. and Wohlleb, C., 1973, **Implementation of Head Start Planned Variation:1970–71.** Part 1 and 2, Cambridge, Mass.: The Haron Institute.

McLaughlin, Milbery and David D. Marsh, 1978, Staff Development and School Change, **Teachers College Record,** v. 80, 69–94.

McGaw, B, 1985, Meta-analysis. In Husen T., and T.N. Postlethwaite T.N., (eds.), International Encyclopedia of Education, Oxford: Pergamon Press.

Morris, L.L. and Fitz-Gibbon, C., 1978, How to Measure Program Implementation. Beverly Hills: Sage.

National Science Foundation, 1978, Case Studies in Science Education, Vol. 1, The Case Reports. Center for Instructional Research and Curriculum Evaluation. Urbana: University of Illinois.

Owen, John M., 1979, User Characteristics Associated with the Implementation of a Curriculum Innovation – An Australian Example. In **Tamir P. (ed) Curriculum Implementation.** Jerusalem: Hebrew University, Science Teaching Center.

Parlett, M. and Hamilton, D., 1977, Evaluation as Illumination: New Approach to Study Innovative Programs. In Hamilton et al (eds): **Beyond the Numbers Game,** Berkeley: McCutchen.

Rogers, Everett M. and Shoemaker, F., 1971, **Communication of Innovations: A Cross Cultural Approach.** New York: The Free Press.

Sabar, N., 1986, Is the Disappointment in the Implementation of a New science Curriculum Justified? A Study Quantifying Descriptive Data. **Journal of Research in Science Teaching.** 23, 475–91.

Scheirer, M.A. and Rezmovic, E.L., 1983, Measuring Degree of Program Implementation: A Methodological Review, **Evaluation Review,** v.7, 599–633.

Scriven, M., 1967, The Methodology of Evaluation, In Tyler, R. (ed) **Perspectives of Curriculum Evaluation,** Chicago: Rand McNally.

Scriven, Michael, 1972, Prose and Cons about Goal Free Evaluation. **Evaluation Comment,** 3, 1–4.

Silberstein, M., 1980, **Planning and Empirical Validation of a Design for Evaluating the Implementation of a Curriculum,** Ph.D. Thesis, Tel Aviv University.

Smith, Louis M., 1978, Science Education in the Alte Schools: A Kind of Case Study. In National Science Foundation: **Case Studies in Science Education.**

Stake, Robert, 1978, The Case Study Method in Social Inquiry. **Educational Researcher,** v. 7, 5–8.

Stiles, Lindley and Beecham, Robinson, 1973, Change in Education, in G. Zaltman and R. Schwartz (eds.) **Perspectives on Social Change,** New York: Wiley Interscience.

Wang, M. et al., 1984, The Utility of Degree of Implementation Measures, **Curriculum Inquiry,** v. 14, 249–86.

Welch, W. and Ward, W., 1975, Evaluation of USMES Implementation Projects (Mimero) Minnesota Research and Evaluation Project. Minneapolis: University of Minnesota.

Yoyole, A. 1977. Observation Techniques in Lewy, A. (ed.) **Handbook of Curriculum Evaluation,** New York: Longmans.

Chapter Three

Quantitative and Qualitative Paradigms

The distinction "quantitative versus qualitative", refers to types of data used and ways data are summarized within the framework of an evaluation study. Researchers frequently refer to the terms quantitative and qualitative as the most important descripitons of two major study paradigms. For a very long time these two paradigms represented conflicting views about appropriate ways of evaluating curricula. Recently, however we witness a kind of reconcilliation between these two approaches. In Lewy's definition of curriculum evaluation (see fig. 3.1) the qualitative/quantitative dichotomy was included in a list of six salient features characterizing evaluation studies, such as criteria, the purpose of the study, etc. (Davis 1981).

Evaluation is the provision of information at the [pre-development / development / dissemination / installed programme] [stage of curriculum development]

concerning [aims / contexts / materials / transactions / outcomes]

from the point of view of [eliciting processes / yielding outcomes / meeting standards / avoiding problems]

on the basis of [measurement / description / observation]

summarized in a [qualitative / quantitative] way

judgment

for the sake of making descisions about [modifying / selecting elements of / qualifying the use of / accepting or rejecting] the program

Figure 3.1 Mapping sentence of curriculum evaluation
Davis (1981) modified from Lewy (1977)

Dorr-Bremme (1985) prefers to use the terms naturalistic versus "experimental or psychometric oriented" evaluation. He characterizes research paradigms by clusters of attributes which also contain the quantitative-qualitative distinction. According to his view the experimental or psychometric oriented evaluation is associated.

> "...with such terms as quantitative research, scientific research, hard data, variables, hypothesis testing, and statistics. In contrast, naturalistic inquiry is commonly associated with such phrases as qualitative research, ethnographic research, field work, case-study method, soft data, ecological, descriptive and interpretive."(p.3490)

To these two clusters of descriptive characterization one may add the "structured versus unstructured" distinction as suggested by Seashore-Louis (1982). She claimed that the term "structured" is used to mean formally and systematically organized while "unstructured" refers to the lack of systematic organization, i.e. loose, open, and free.

Guba and Lincoln (1985) strive to provide a conceptually clear distinction between the two dominant paradigms of disciplined enquiry and claim that it is a mistake to view the terms "qualitative", "case study" and "naturalistic" as logically linked together, despite the facts that in research practice they appear to be clustered together. Guba and Lincoln refer to the two distinctive paradigms as Naturalistic and Rationalistic, and point out the following five axiomatic differences between them.

1. The nature of reality – The rationalists believe that there exists a single reality which can be broken down into component elements, or variables and each of them can be studied independently of others. The naturalists believe that there exist multiple realities, each of them representing different meanings and interpretations attributed to a certain phenomena by various persons. Reality is intangible and should be studied in a holistic fashion.
2. Inquirer-respondent relationship – Rationalists believe that the researcher should take an objective stance toward the target of his or her study, and should avoid intensive interaction with respondents or suppliers of data. Naturalists encourage interaction with the resopndents and suggest that respondents' statements should be interpreted in the light of the intent

attributed to them by the respondent.

3. The Nature of truth statements – Rationalists strive to produce generalizations, which have context-free enduring value. Naturalists believe that human behavior is never context-free and therefore there is no possibility of producing generalizations. The evaluator should strive to describe a particular case, and produce "working hypotheses" which are transferable from one context to another "depending on the similarity of contexts".

4. Causality – Rationalists believe that their enquiry explores cause-effect relationships. They adhere to Hume's if-then formulation of such relationships. In contrast, naturalists believe that the best the researcher can do is to present "plausible patterns of influence".

5. Relation to values – The rationalists believe that the enquiry process is objective and value-free, while naturalists claim that selecting a research problem, rooting that problem into some theoretical framework, and preferring one methodology to another – let alone the interpretation of results – all entail identification with certain values.

In spite of the convincing arguments of Guba and Lincoln (1985), practitioners tend to use the distinction quantitative-qualitative for describing the two commonly used evaluation paradigms, probably because they require differential trainings and differential research skills. Therefore our adoption of these terms represents a matter of convenience only, and by no means does it reflect a different conceptual view than that described above.

As already indicated, the terms quantitative and qualitative are generic labels. They refer to some well-defined aspects of evaluation studies, but in and of themselves they entail nothing about other aspects of study (Dorr-Bremme 1985). Within both of these categories one may find a variety of study types. To illustrate the unique features of the two paradigms some instances of curriculum evaluation studies will be described.

The Quantitative-Qualitative Continuum

The principles described above refer to archetypes of the two study paradigms. But a great many evaluation studies cannot be easily described as fully following one or other of the paradigms. They represent some in-between position in relationship to the two. Seashore-Louis (1982) suggests that one should not regard these two paradigms as a dichotomous division of actual studies, but rather as poles of a continuum on which a variety of actual studies can be located.

In an attempt to provide a systematic framework, she describes three dimensions of evaluation studies, wich are related to the quantitative – qualitative continuum. She adds to the routinely-used data collection and data summary dimensions a third one, i.e. the data base formation dimension (see fig. 3.2).

This third dimension represents an intervening step between data collection and data summary, and it becomes necessary due to the fact that the original data collected in evaluation studies, quite frequently, are transformed before they are submitted to analysis. Thus, for example, narrative descriptions of classroom events may be classified according to prespecified categories, and thus converted into a quantitative data base for the purpose of statistical analysis. Conversely, open-ended responses to questionnaire items in a survey-type study can be summarized in qualitative way.

As specified in figure 3.2, the data collection technique may be highly structured, as is the case with census data, or highly unstructured, as in ethnographic observations. Between these two ends of the continuum one may find a variety of in-between techniques. Likewise along the data-base and the data summary dimensions one may find a variety of categories which represent transitions between the two archetype poles.

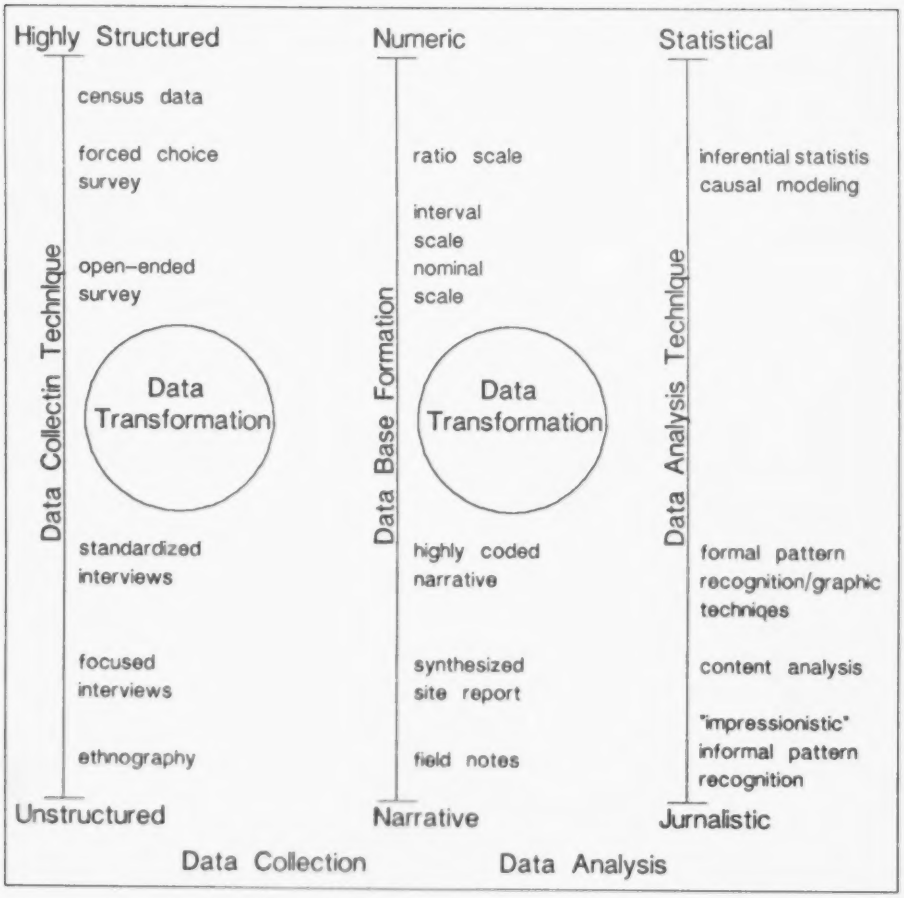


Figure 3.2: Dimensions of variation in multisite/multimethod studies

Examples of Quantitative and Qualitative Studies

To illustrate the differences between these two study types, excerpts from, and abstracts of quantitative and qualitative studies will be presented here.

Quantitative Studies

Quantitative studies are characterized by Dorr-Bremme (1985) as either experimental or psychometrically oriented. One may also add to these study types surveys and quasi-experimental studies.

In the history of curriculum evaluation, studies following this paradigm were dominant for many years. In the 1960's and early 1970's this paradigm was the sole legitimate one (Weiss 1972). Its dominance is well demonstrated by a survey conducted in 1976. Forty-eight well known experts were asked to rate the importance of various methodological skills for people working in the field of evaluation (Anderson and Ball 1978). The highest rating was given to statistics and experimental design. At the bottom of the list of 32 skills were content analysis, case study methodology, and job analysis.

The abstacts of studies presented here illustrate two features of the quantitative evaluation mode: experimental manipulation and the psychometric method of developing measurement scales.

Example 1: Experimental manipulation

A classic example of experimental manipulation used to evaluate a planned curricular innovation is provided by Gall et al (1978). The aim of the study was to examine the merits of curriculum materials which contain a large proportion of questions requiring higher-order mental operations. For the purpose of the experiment they constructed a short-duration learning unit consisting of five lessons. The materials described for all five single period lessons contained eight major questions for each period. Two different versions of the instructional material were prepared. In one version, 75 percent of the questions required higher-order mental operations (such as applying principles, analyzing situations, etc.) and 25 percent of the questions required lower-order mental operations (such as recall of information, etc.). In the other version, the proportion of question types was exactly inverted. An experiment was carried out by

selecting two samples of classes from the same population. One of the samples used the higher mental operation materials and the other sample used the second version of the materials. Teachers were randomly assigned to the groups, and they were carefully trained to adhere to the prescribed instructions. The evaluators referred to the materials as "semi-programmed" instructional materials, since they strictly defined the content of instruction and the nature of the questions. But unlike a situation of programmed instruction, they provided some freedom for the teacher to respond to emerging questions. Despite the utilization of prescribed instructional materials and exactly-worded questions, the natural flow of the lessons was not disturbed, and observers indicated that they did not notice any unusual or rigidly structured pattern of classroom occurences.

In order to keep the experiment under control and to eliminate the interference of contaminating effects, such as familiarity with the topic taught, or differential impact of teachers, the evaluators introduced some control devices. First, they selected a topic which was not dealt with in the school curriculum and therefore they could assume that all students were equally unfamiliar with the course content. Secondly, the teaching was carried out by visiting student-teachers who did not have previous contact with the learners. They were randomly assigned to the classes. By this arrangement it was possible to avoid the effect of previously established teacher-student relationships.

The results of the study, whatever they were, are not our concern in this context. The example mentioned here illustrates the ingenuity needed for isolating the independent variable, i.e. the proportion of higher-order questions, and for keeping constant other variables, such as familiarity with content, or differential student-teacher relationships.

An artificial situation had to be created which differed from the regular class situation in at least three aspects:

1. The content of the course was not a part of the regular curriculum.
2. The questions to be posed were strictly prescribed and exactly worded by the planners of the experiment.
3. The teachers did not have any previous contact with the learners.

It should be noted that precisely these non-standard, artificial teaching-learning circumstances guarantee the validity of the experimental results. But due to the deviation from routine classroom situation it is not surprising that some educators expressed doubts about the generalizability of the findings to routine classroom practice.

Example 2: A psychometric measure of implementation.
Unlike the experimental design type evaluation described above, the evaluation study of Leinhardt (1977) was carried out in an actual teaching setting with no direct intervention in the teaching-learning process on behalf of the evaluator. The evaluator observed classroom transactions, collected and analyzed quantitative data, but did not carry out any manipulations which directly touched upon the classroom activities. The unique feature of the study was the way the evaluator developed a scale for measuring the program implementation in various classes. The development of the measure of implementation contained several steps. The researcher collected a large quantity of data which intuitively seemed to be related to the implementation of the program, and then used statistical procedures to reduce the data into an interpretable set of variables. As the first step of data reduction, she eliminated measures which had no variance. Next, she selected one variable only from sets of variables which exhibited a relatively high level of intercorrelations. And as a third step, she looked for ways to combine variables into clusters. The researcher decided not to use factor analysis for this purpose, because factor patterns are often illogical, that is, the variables that go together mathematically do not do so logically, and often do not replicate. She preferred to use for this purpose the Cooley and Lohnes (1976) model. The basic approach of this model is selecting

> "...variables which can logically be assigned to each of the major constructs defined, and then to combine the variables within each construct. Combining the variables is accomplished by factoring within the construct and summing using the factor loading, or by transforming the raw data to standard scores and summing using equal weights. This approach is most useful if the data are initially collected with the model in mind rather than if the data are artificially molded to the conceptual model" (p. 285)

Some clusters of variables identified by this technique are presented in Table 3.1

Table 3.1 – Clusters of variables

Instructional Behaviors
Active responses sought
Modeling of response
Focusing attention
Contingent praise
General praise
Clear tutoring
Class management
Backward chains
Refers to earlier information
Solicits opinions
Refers to earlier success
Opportunity
The number of adults travelling (moving) in class
Availability of trained alternates
Percentage of children present
Time estimate
Minutes a day spent using individualized programs
Autonomy
Children decide what unit to study
Children request special tutoring
Children starting to work as soon as they come in
Peer tutoring
Assignments
Percentage of unique assignments

After deriving measures of implementation for each class, the evaluator examined the contribution of program implementation to the program outcomes. This was accomplished through highly sophisticated statistical analysis known as "commonality analysis" (Mayeske 1970).

The details presented above constitute a highly simplified version of the actual statistical manipulation carried out by the evaluator. Though the evaluator has not interfered with the routine class operations, nevertheless, the variables she developed, the measures and the criteria she employed, are all very different elements from those which concern whoever particpated in the observed teaching-learning situation.

The strength of this approach is in the exact quantification of the measures, and its weakness is in the focus on a relatively few selected aspects of a complex phenomenon, disregarding all other aspects which were not captured by the prespecified data collection scheme.

Qualitative Studies

The term "qualitative" is used for describing a great variety of study types. Their common feature is that the summaries are not based on numerical manipulations of quantified data, and are not presented in numerical terms. Usually, the data collected neither have numerical character nor do they become transformed into numerically coded records. But there are qualitative studies which also utilize numerical data. Among the varieties of qualitative evaluation studies one may find the following categories:

Evaluative case study – Evaluative case studies may focus on a single case or a collection of cases. The evaluator collects a variety of data type and studies the case in depth with the aim of helping the decision maker to judge the merit of an entity (program, institute, teachers, etc.) (Stenhouse 1985).

Intrinsic evaluation studies – Program evaluation studies of this type are based on analytical examination of texts, documents, etc., focusing on the merits of the goals, the consistency of its component elements, and their fit to prespecified standards (Eraut 1985).

Judicial evaluation studies – Judicial evaluation is a method which combines procedures both from court proceedings and administrative hearings. Concepts such as fact-finding, human testimony, cross examination, case preparation, evidentiary rules and procedures of structured deliberation were borrowed from the legal system and modified for evaluation purposes (Wolf

1980).

Curriculum criticism – Adherents of the curriculum criticism approach claim that curriculum can be treated as a literary object and that criticism can focus on the object itself rather than on circumstances which are related to its development, utilization, or effects. The critic's statements are not derived from data collected with the aim of examining the quality of the program, but rather from the adoption of a methodology entailing his own critical principles and assumptions (Willis 1978). The curriculum critic has to describe the object of his criticism and to make judgements about its qualities. Since the description of an object requires the collection of an infinite number of details, the evaluator has to select aspects or points of view which constitute the focus of his criticism.

Photographic evaluation – Photographic evaluation strives to record visual information which may serve as a basis for describing programs and to evaluate their worth. The method is rooted in the concept of **cinema verité** which distinguished between documentary photography, desgined to prove a thesis, and naturalistic observation. **Cinema verite** intended to record what happened normally and spontaneously. However for evaluation puroposes, photography should serve as a recording instrument rather than as a documentary device (Templin 1979).

Scientific Versus Artistic Approach

Eisner (1981) distinguishes between two approaches of qualitative evaluation: scientific and artistic. He also analyses the differences between these two approaches. Some of his distinctions are:

1. The forms of representation – Scientific work employs formal statements and avoids metaphorical language. Artistic language has no codified language.
2. The role of form – Science uses a standardized style of reporting. For instance, an academic article will usually identify a problem, review the literature, ect. An article reporting on some aspect of the artistic field would tend to be far more idiosyncratic; standardization in this context would be counter-productive.

3. The criteria for appraisal – Scientific research looks for consensually approved criteria. The artistic approach does not need formalized evidence.
4. Points of focus – The scientific study focuses on the manifest behavior, the artistic one uses the manifest behavior as a cue for making inferences about the inobservable.
5. The sources of data – In the artistic approach the major instrument for data generation is the investigator himself, in the scientific approach the investigator relies on some standardized rules for collecting data.

In the subsequent section abstracts of, and excerpts from qualitative studies will be presented. The first two represent the approach which was referred to by Eisner as scientific, and the third represents the artistic approach.

Example 1. Scientific approach to qualitative research: Thirteen large scale case studies.
One of the most comprehensive qualitative studies in the field of curriculum evaluation dealt with the status of pre-college science education in the United States (Stake and Easly 1978).
The researchers defined the broad aims of the study as the answering of the following questions:

- How is science being taught today?
- What are the current conceptualizations of science in the classroom?
- What are the current encroachments upon the science curriculum?

These questions constituted the starter list of a variety of issues. A fuller list of specific evaluation focuses emerged in the course of the study (see Table 3.2).

Table 3.2: List of evaluation focuses.

Different Conceptualizations of Science	**Science Instruction**
Science as inquiry	Appropriateness of teacher-preparation
Science as explanation	Organization for teaching, class periods
Science vs. technology	Testing, assessment, teaching for the test
Social science vs. social studies	Science clubs, competition, honors
	"Driving-force" persons, "Mr. science"
	Changing roles for teachers
Perception of Conflict between Science and Culture	Changes in School or Community that may affect science (including mathematics and social sciences) curricula
Science and Religion	Diminishing budgets for education
Social science and cultural taboos	
Moral issues and science teaching	
Place of Science in the Curriculum	Emphasis on the basic skills
Core subjects and electives	Emphasis on bilingual programs
Preparation for college	Advisory roles of teachers and administrators
Vocational relevance of science	Increasing role of parents and citizens
Integration of subjects	Desegregation actions

The study was carried out by thirteen teams each of which accepted responsibility for conducting a case study in a particular school district.

The teams were quite autonomous and the common features of the studies touched on some aspects of data collection. All teams carried out observations in the class, ran interviews with significant persons at the site, and invited a visiting team to stay at the site for several days to observe relevant phenomena.

The validity of the studies was established on the basis of comparing information derived from various sources. In technical terms this process is referred to as triangulation.

The field observers sought out information from different positions, roles, experiences, attitudes, and goals in order to check the perceived constancy of a phenomenon. The observers themselves observed, interviewed, and analyzed documents. Their findings were reviewed by visiting teams, site coordinators, and on-site educators. All provided additional views as well as confirmations of refutations or particulars.

Writing each case study remained the responsibility of one person, the field observer. The utilization of data collected in the field is illustrated by an excerpt from a case study.

An excerpt from a case study of school science

Several informants noted a general lessening of interest in science as children moved through the elementary grades.

> "First graders love science. It is exciting and different. By the time they get to junior high some are saying, "oh no, not science!" When first and second graders get excited about science – magnets are fun you know – they squeal with delight. We somehow manage to take out that fun as we go along. I don't know how we do it."

Science in general has lost its appeal. The science fairs, so "big" in the sixties, have been reduced to a sometime thing for most schools. Several experts observed that the parents' science fairs of the sixties were a mixed blessing. The fairs at once created parental interest in science education and "terrific competition among the parents for recognition and prizes warded to (their and) their children's entries." Teachers gave broad hints at what "works" in teaching science to elementary age children.

"The things that turn on third graders are things which they have seen or heard about but never really understood."

"If something does not appear to the child to be related directly to his life then it had better be amazing or fascinating. Teacher education is really short of giving us ideas that work."

"More than anything else, children love experiments. Even watching me do one is better than reading. Doing them is best of course, but there is very little they can really do. Right now we are creating rust. When they read ferrous oxide it means little. But when they see it form, it sticks. This is where the equipment and space become crucial. Faces light up for activities, go blank for reading about science."

"The things that work are not surprising. When children are interested, are active and involved, classes pass nicely."

One of the things that "works" is to avoid assigning reading in science. A similar observation was made by teachers in arithmetic and social studies in all elementary school grades. Reading, mathematics (and spelling) come first. Science and social studies come second in both teacher preparation time and actual teaching time.

Example 2. Scientific approach to qualitative research: Observing a single pupil

McDermott et al (1984) used qualitative research methodology in a study designed to examine differences among pupils in doing their homework assignments. Each member of the team visited a single child at his/her home day after day for a period of a few weeks and observed the child doing the homework. The authors decided to present a thorough description of two pupils preparing their homework.

Both of them were working class children studying at the same grade levels in schools of comparable quality. Both prepared homework under the supervision of one of the parents. The researchers found that the differences between the children in doing their homework can hardly be described in terms of achievement related variables such as motivation, prerequisite variables, school attitude, academic aspiration etc. Instead, they identified three variables which, in their opinion, best described the differences between the two pupils: ritualism, working space, and continuity.

For one of the children, preparing homework was associated with a high level of ritualism. The pencil had to be sharpened, the sharpener, the ruler, the dictionary, the working book etc. had to be arranged in a neat order. He did the work in a quiet room, where he was not supposed to be disturbed, and brothers and sisters were not allowed to play in that room while he studied. The process of the work was nevertheless frequently interrupted, and at each occasion the rite of arranging the accessories on the desk had to be repeated and the work started again. The researchers characterized the sequence of activities as "starting, interruption and starting again".

In contrast, the other child did his homework as an integral part of the afternoon activities. He was together with other members of the family in the same room. The mother watched TV, ironed clothes or took care of other children in the room. Whenever an interruption occurred the child quietly returned to his work and continued to do the assignments. The researcher characterized the

sequence of activities as "starting, interruption and continuing".

The study described above illustrates the scientific value of observing a single child and collecting information about his behavior. The observation proved useful because it succeeded in identifying meaningful differences in the way children deal with homework, and the findings of the study may serve as a basis for adapting assignments to the working pattern of the child.

Example 3. The artistic approach in qualitative evaluation

An evaluation study carried out by Jenkins (1978) describes a week-long residential course at the London Business School offered to business executive members. The evaluator participated in one of the courses and his criticism reflects the insider's personal perspective. Excerpts from the report illustrate the way the evaluator described various aspects of the program.
The first impressions of the participant observation are described as follows:

> "The visitor to London Business School has an immediate impression of the theatricality of its setting. The elegant Nash facade overlooks Regent's Park at a discreet distance from London Zoo and its Snowdon bird house. Its white pepper-pot turrets crown a columned facade. Only later, viewing the flat cardboard cutout ballustrades from inside a window, does the visitor realize that the facade is a set piece of stage decor for LBS street theater. It makes a briskly functional building. (p. 345)

The evaluator provides details about the program and about people who developed the program.

> "Another protagonist in our drama is P.B., director of a computer-assisted learning unit at LBS. P. is young, watchful, nerve-wracked, and angular as a hairpin, but quite exceptionally bright and determined to make his mark through promoting what might, not unfairly, be described as a "pedagogical invention." (p.346)

> P's own "invention" is a package that applies computers to management education in two ways. The first is a simulation excercise, usually known to man and computer alike as EX-SIM. This is a simulation of a "decision-making environment" inhabited by competing managers, trade union officials, and others. In the Skills of Bargaining course, EX-SIM was an exercise called Ball Bearings Ltd. A succession of decisions in time (called "rounds", as in professional wrestling) allows the managers to live with the consequences of their actions. The second part of the invention, which is a prime interest in this report, is a feedback process unsuprisingly called FDBACK. The purpose of feedback is to give managers access to their own and other people's percep-

tions of problems and events. The emphasis is on explaining concepts.

(p. 346)

The program is characterized as employing a mixed pedagogy.

"There is, however, an air of mystique about FDBACK. like most games, it is learned in play more readily than through explanation. Those who understand it technically and argue about such things as the relative merits of product-moment and rank-order correlation may fail to have any real sense of its educational potential; those who find it exciting in use may be puzzled by its epistemoiogical status, particularly as claims on its behalf differ wildly. At times it is presented as a "computerized scratch pad," almost a way of allowing managers to produce collaborative doodles on their mental life. At other times it is sold hard as an X-ray photograph of particular minds at work, using Kelly's theory of personal constructs as the plate. Paradoxically, this may not matter. Like the Tarot pack, its significance depends upon interpretation and use..." (p.349)

From Disdain to Détante

For decades the two paradigms of research, the qualitative and the quantitative ones, have represented conflicting views, and the gulf between them has constantly deepened. This is so despite the fact that for more than a quarter of century Zetterberg (1965) advanced the hypothesis that quantitative and qualitative approaches have distinct roles in the advancement of disciplinary knowledge and that both of them are essential. Qualitative field work can help to illuminate the theoretical and measurement issues in new areas of study, while quantitative methods are appropriate for testing the hypothesis derived from rigorous field methods.

The gulf has been precipitated due to the advancement of sophisticated mathematical models of data analysis, on the one hand, and the diversification of qualitative research and evaluation models, as described above, on the other. In the 1970's experts jealously argued for the priority of one or another research paradigm of evaluation. Thus, for example, Stake (1976) indicates preference for the qualitative paradigm of evaluation. He writes:

"A major purpose of an evaluation study is to alleviate a problem. I put higher priority on helping solve a problem of a particular program than on gaining generalizable understanding in several. And I have not found an experimentalist approach very useful for problem solving". (p.20)

Stake claims that evaluators have to produce findings which are comprehensible for the decision maker, and therefore he recommends providing comprehensive portrayals of what is happening, and to thereby produce vicarious experience. The abstract data of the researcher are not well understood by decision makers in education, and therefore they are not adequately used. On the other hand, Boruch (1982) believes that innovations should not be installed prior to their tryout and evaluation, and strives to prove that the randomized, experimental design is the most valid method to demonstrate the merits of any innovation.

> "We recommend that pilot tests be undertaken before new programs or new variations are adopted, and that the introduction of new programs be staged so that good designs can be exploited. Further, we recommend that higher quality evaluation designs, especially randomized experiments, be authorized, explicitly in law, for testing new programs, new variations of existing programs, and new program components". (p. 1)

Boruch rejects the claims that it may be difficult to carry out randomized experiments in field settings and lists numerous examples of carrying out such evaluation studies. He concludes as unjustified claims that it is impossible to assign individuals (or other units) randomly to programs for the sake of fair estimates of program effects.

Nevertheless, towards the end of the 1970's attempts were made to reduce the conflict between the two camps and to create some armistice between them. Rist (1977) refers to these attempts as a movement from disdain to détente. One can also read about reconciliation and the mixing of the two methods (Jick 1979).

Why is the Mixing of Methods Recommended?

Jick (1979) attributes the quest for mixing multiple research methods in social sciences, to the impact of the seminal article of Campbell and Fiske (1959). That article first introduced the concept of "convergent validity," which later became referred to as triangulation. The metaphor triangulation is borrowed from navigation and military strategy which use multiple reference points to locate an object's exact position. According to this conception qualitative and quantitative methods complement each other rather than contradict each other. This view goes beyond what was suggested by Zetterberg (1965) that these two methods have a distinct function and are appropriate at different phases of generating knowledge. In wake of the Campbell and Fiske (1959) article there has been a demand for the simultaneous use of different research paradigms for responding to a single research question. They express the view that convergence of findings obtained through two methods enhances the belief that the results are valid, and that they are not a methodological artifact. A thorough presentation of cross validation procedures of triangulation techniques is provided by Denzin (1978).

Using multiple methods for evaluating a particular educational program would involve conducting interviews with persons who are "stakeholders" in the program, conducting an intrinsic evaluation study of the materials used, measuring the implementation of the program, collecting data about mastery of educational objectives, etc. and integrating the results obtained from various sources, and processed through various methods. Jick (1979) added that triangulation does more than merely provide evidence about validity. It also captures a more complete, holistic, and contextual portrayal of the target of the evaluation.

> "It is there that qualitative methods, in particular, can play an especially prominent role by eliciting data and suggesting conclusions by which other methods would be blind. Elements of the context are illuminated. In this sense, triangulation may be used not only to examine the same phenomenon from multiple perspectives but also to enrich our understanding by allowing for new or deeper dimensions to emerge.
>
> (Jick, 1979 p.604)

Also McGrath et al (1982) emphasize the complementary nature of the two research paradigms. They claim that each research paradigm suffers from certain flaws. There is no ideal research paradigm.

> Strategies that yield precicion, such as laboratory experiment, for example, also yield narrow scope and low realism. Measures that offer one kind of rigor, such as paired comparison judgements, often are highly reactive. Data analysis methods, that give "strong" answers, such as multiple regression analysis, do so by imposing many and strong assumptions; hence the methods themselves strongly determine the responses. (p.108)

In the face of the methodological weakness of all research paradigms the response to a particular question cannot be fully given through a study using one particular methodology. Knowledge acquisition is an accrual process. "A single study is never the first step in the knowledge accrual process, and no study is ever the final step (McGrath et al 1982, p.107)"

The desirability of examining each particular phenomenon by a diversity of methods does not mean that each individual researcher has to be an expert in all possible techniques or methods of inquiry. It is sufficient if each individual researcher and evaluator appreciates the potential value of all methodologies and is also aware of their limitations. Knowledge accrues through the joint effort of the whole community of researchers in which expertise in a variety of methodologies and study paradigms is usually well balanced. In contemporary research enterprise the role of the stand-alone researcher, which was so dominant in educational research several decades ago (Baker 1967) is gradually diminshing. Research is more often being carried out by teams, and issues of crucial theoretical or social significance are, as a rule, examined by several teams. Consequently there is no technical difficulty in solving the problem of employing multimethod approaches to the study of a particular issue. But to do this successfully requires more than a passive tolerance of "other" methods. Deténte does not provide a solution. There is a great need for conviction, persuasion and effort.

References

Anderson, S. and Ball, S., 1978, **The Profession and Practice of Program Evaluation**, San Fransisco, Jossey Bass.

Baker, Frank B., 1967, Experimental Design Considerations Associated with Large Scale Research Projects, In Seventh Annual Phi Delta Kappa Symposium on Educational Research: **Improving Experimental Design and Statistical Analysis**. Chicago, Rand McNally.

Boruch, R., 1982, Experimental Tests in Education: Recommendations from the Holtzman Report. **The American Statistican**, 36, 1–14.

Campbell, D.T. and Fiske, D.W., 1959, Convergent and Descriminant Validation by the Multitrait – Multimethod Matrix **Physchological Bulletin** 56, 81–105.

Cooley, W.W. and Lohnes, P.R., 1976, **Evaluative Inquiry in Education** New York, Irvigston Pbl.

Davis, Ed., 1981, **Teachers as Curriculum Developers** Sydney, Allen and Unwin,.

Denzin, Norman K., 1978, **The Research Act**, 2nd ed. New York, McGraw-Hill.

Dorr-Bremme, D.W., 1985, Naturalistic Evaluation, In Husen T. and Postlewaite, T.N. (eds.) **International Encyclopedia of Education**

Eisner, E.W., 1981, On the Differences between Scientific and Artistic Approaches to Qualitative Research. **Educational Researcher** 10(4), 5–9.

Eraut, M., 1985, Intrinsic Evaluation, In Husen, T. and Postlewaite, T.N. (eds.) **International Encyclopaedia of Education**.

Gall D.M., Ward, B.A. Berliner, D.C. et al, 1978, Effects of Questioning Techniques and Recitation on Student Learning. **American Educational Research Journal**, 15, 175–200.

Guba, E.G. and Lincoln, Y.S., 1985, Naturalistic and Rationalistic Enquiry, In Husen, T. and Postlethwaite, T.N. (eds.) **International Encyclopaedia of Education.**

Husen, T. and Postlethwaite, T.N. (eds.), 1985, **International Encyclopaedia of Education: Research and Studies.** Oxford, Program Press.

Jenkins, David, 1978, Business as usual: The "Skills of Bargaining" course in LBS, In Willis, G.(ed.) **Qualitative Evaluation**, San Francisco, McCutchin.

Jick, Todd D., 1978, Mixing Qualitive and Quantitive Methods: Triangulation in Action **Administrative Science Quarterly** v. 24, 602--611.

Leinhardt, Gaea, 1977, Program Evaluation: An Empirical Study of Individualized Instruction. **American Educational Research Journal** v.14, 277–293.

McGrath, J.E. Martin, J., and Kulka, R.A., 1982, **Judgement Calls in Research,** Beverly Hills, Sage Publ.

McDermott, R.P., Goldman, S.V., and Varenne, H., 1984, When School Goes Home: Some Problems in the Organization of Homework. **Teachers College Record,** 85, 391–409.

Mayeske, G.W., 1970, Teacher Attributes and School Achievement. In Office of Education (ed.) **Do Teachers Make a Difference?** Washington D.C.: Government Printing Office.

Rist, R., 1977, On the Relations among Educational Research Paradigms: From Disdain to Detente. **Anthropology and Education Quarterly,** 8 (May): 42–49.

Seashore-Louis, Karen, 1982, Multisite/Multimethod Studies: An Introduction. **American Behavioral Scientist**, 26, 6–22.

Stake, R., 1976, Theoretical Statement of Responsive Evaluation. **Studies in Educational Evaluation,** 2, 19–22

Stake, R., and Easly, J., 1978, **Case Studies in Science Education 2. vols.** Urbana, University of Illinois, CIRCE

Stenhouse, L., 1985, Case Study Methods, In. Husen, T. and Postlethwaite, T.N. (eds.): **International Encyclopaedia of Education**.

Templin, P., 1979, Photography in Evaluation **Newsletter: Northwest Regional Laboratory** v.2(1), 1–2

Weiss, C., 1972, **Evaluation Research: Methods of Assessing Program Effectiveness.** New York, Prentice Hall

Willis, G., 1978, **Qualitative Evaluation**. San Francisco, McCutchin

Wolf, R.L., 1980, **Judicial Evaluation: Past, Present and Future Application.** Washington D.C. National Institute of Education

Zetterberg, H., 1965, **On Theory and Verification in Sociology**. Tottowa, N.J., Bedminster.

Chapter Four

Utilizing Evaluation Results

Evaluation studies emerge in order to satisfy the information needs of decision makers. This is explicitly stated in Alkin's (1970) definition of evaluation:

> Evaluation has been defined as the process of ascertaining the decision areas of concern, selecting appropriate information and collecting and analysing information in order to report a summary of data useful for decision makers in selecting among alternatives.
>
> (p.2)

According to this definition evaluation is always related to some action of decision making. In selecting among perceived alternatives, the role of the evaluator exceeds the boundaries of assisting the decision makers. His/her role is also to call attention to the existence of alternatives – even if the decision maker him/herself is not aware of them (Lewy 1977). In this respect, evaluation differs from traditional scientific research which is more concerned with revealing "truth" and accumulating knowledge, than with the immediate utility of the knowledge produced. Evaluation deals with concrete situations, it strives to appraise the worth of a particular entity (such as a person, program, process, product) rather than produce generalizations. Alkin's (1970) early definition is quite restrictive. It specifies only one type of utilization: selecting alternatives. In the context of curriculum evaluation, Davis (1981) specifies a broader range of possible decision types such as modifying the program, selecting certain component elements of a program, qualifying the use of the program or acceptance versus rejection of it.

Patton (1985) provides a broader definition of utilization:

> ...there are a variety of ways in which evaluation processes and findings are utilized. Evaluation can directly influence major, specific decisions. Evaluation can be used to make major adjustments in programs. Decision makers can, and do, use evaluation for reducing uncertainty, enlarge their options, increase control over program activities, and increase their sophistication about program processes. Sometimes evaluations have more of a conceptual impact, that is, they influence how stakeholders think about a program, rather than an instrumental impact, that is, evaluation utilization manifisted in concrete actions and explicit decisions.

(p.1749)

One may observe a trend towards a broadening of the scope of utilization manifestations. Thus, for example, Cousins and Leithwood (1986) claim that recently

> even a more basic conception of evaluation use has been descried: that the more psychological processing of evaluation results constitute use, without necessarily informing decisions, dictating actions, or changing thinking.
>
> (p.332)

According to this definition the processing of evaluation results in itself constitutes use. Moreover, administrators frequently claim that being aware of the existence of an evaluation study, even without considering its results, has an impact on the behavior of various stakeholders in the program being studied.

Alkin himself revised his earlier definition of utilization, and instead of referring to selection from alternatives as the single instance of utilization, he provides a broader list of utilization categories: making decisions, substantiating previous decisions of actions, and establishing or altering attitudes. (Alkin et al 1979)

In an attempt to summarize trends in utilization research in the 1970's a distinction is made between two alternative approaches to defining evaluation utilization.

> The mainstream perspective employs a rather narrow definition of utilization that stresses the rapid decisive impact of the evaluation on major program activities and decisions. The alternative perspective softens these rather stringent requirements, broadening the concept of utilization to include impacts which are more subtle and impacts which are not felt for a period of time.
>
> (Alkin et al 1979, p.223)

For the purpose of conducting empirical studies on evaluation utilization Alkin et al (1979) prefers to employ a broad definition of the term, but not as broad as the previously quoted definition of Cousins and Leithwood (1986).

To arrive at a definition that is flexible enough to encompass the variety of the types of utilization which occur, without also capturing inconseqential spin-off effects.

Alkin et al (1979) suggest isolating essential components of utilization. Three such components should be considered: the type of information, the user(s), and the way the information is used.

The type of information

Most evaluation information contains quantitative summaries of data collected in "field" settings, and these are conveyed to users in the form of written reports. However, this is not the only legitimate kind of information and type of presentation. The information need not necessarily contain quantitative summaries of data. It may also consist of an account of interviews, impressions gained at a site visit, artistic description of a situation, etc. The form of communication, too, need not necessarily be a written report. It may also be a conversation with decision makers, an oral presentation of the findings, and exhibition of products, a photographic presentation of significant occurrences, or a film or videotape record of events.

The user(s)

The second component of evaluation described by Alkin et al (1979) is the user. They distinguished between admissible and non-admissible users.

> Of course not every user or use is truly admissible. Suppose, for example, an evaluator used a particularly novel analytical procedure, which was reported in a subsequent journal article. Use of this procedure by other evaluators who had read the article would, in our mind, hardly count as an instance of "evaluation utilization". This is an extreme case but it does draw our attention to the need to limit our definition of admissible users and uses so that such tangential uses are not counted as instances of utilization.
>
> (p.227)

Furthermore, there are two types of admissible users. There are those who commissioned the study and have authority to supervise the evaluation activities. The evaluation literature refers to persons in this category as clients. The second group of admissible users are those who are sanctioned by the clients.

In the context of curriculum evaluation a project director, or organizations that initiated a project and assumed responsibility to finance it may be the direct clients of an evaluation study. Sanctioned users may be school teachers, or an organized group of parents. Though they are not authorized to oversee the evaluation, they may use evaluation results for improving the school program.

Alkin et al (1979) emphasize the authority structure of various users' groups and attribute greater significance to utilization on behalf of sanctioned users. In contrast, Lewy (1981) emphasizes the advantages of simultaneously communicating the evaluation results to various stakeholder groups of a program. Serving different groups of stakeholders at the same moment may enhance utilization. This is so because groups representing different interests are less likely to create a coalition for disregarding information than members of the same interest group.

The way of using information

The "idealized" mode of use is a decision made or an action taken on the basis of a particular information input, and only on the basis of that information. In reality, however, one seldom encounters situations where a decision is made after considering only one particular course of information. Quite commonly, one source of information is matched with information from other sources. In the actual process of decision making any one of the information sources may be dominant, but it will not remain the single determinant of the decision. Evaluation information may be one of various factors affecting action. The decision-maker may combine the evaluation information with information derived from contextual factors, political considerations, personal experience and so forth.

There are cases when the evaluation information has only a cumulative influence. This may occur when, during the life cycle of a single program evaluation, studies are carried out at different periods of time. Consequently findings emerging repeatedly at various periods of the program's life cycle may lead toward action even if the findings of a single set of data have not done so. Cumulative influence may also be experienced if a variety of evaluation studies, dealing with different programs, yield congruent results. In such cases, it may well be that evaluation results produced within the framework of a single study have not directly motivated action, but may provide support for information derived elsewhere, such as from another study on the effectiveness of

another program. The evaluation client engaged in producing and/or operating a series of programs, may use information gathered from examining one particular program for making decisons at a later time, when he/she must make decisions about other programs.

Factors Affecting Utilization

Several attempts have been made to group factors which may promote utilization. Patton (1977) clustered these factors into three groups: organizational factors (referring mainly to the characteristics of the organizations that are expected to use the evaluation results), personal factors (related both to the evaluator and the user), and finally the characteristics of the study. Alkin et al (1979) present a more detailed classification scheme of factors as specified in Table 4.1.

Table 4.1 Properties affecting evaluation utilization

(1)	Pre-existing evaluation boundaries (such as fiscal constraints, or school-community relationships).
(2)	Orientation of the users (expectations from evaluation, preference).
(3)	Evaluator's approach (selecting model, attributing importance to rapport with user).
(4)	Evaluator credibility.
(5)	Organizational factors (the organizational features of the project being evaluated).
(6)	Extraorganizational factors (community, government influences).
(7)	Information content and reporting (format, dialogue, etc.).
(8)	Administration style (initiative, leadership).

Cousins and Leithwood (1986) distinguished between two major categories of factors: evaluation implementation (referring to all process characteristics of conducting the study and to the charcteristics of the evaluation product) and decision and policy setting (referring mainly to the characteristics of the users' organizational setting).

The common feature of these classification attempts is reference to the evaluation product itself, to the process of conducting evaluation, orientation towerd the user and to the characteristics of the users. These four issues will be used as a framework for our dealing with factors affecting utilization.

Characteristics of the Evaluation Product

The quality of the evaluation product has been considered for a long time the single valid and relevant criterion to assess the merit of an evaluation study. Evaluation adopted the classical experimental design pattern and complied with rigorous methodological prescriptions, and this seemed to be sufficient to guarantee the high quality of an evaluation study. Recently, however, a broader variety of criteria are used to evaluate the quality of evaluation products. **Timeliness** – Stufflebeam (1969) claimed that the classical "precision" criteria, such as validity and reliability, are not sufficient to guarantee the high quality of an evaluation study and suggested several other criteria, among them **"timeliness"**, which he defined as the availability of the information at the time it is needed for decision making. The Evaluation Standards, too, address themselves to the issue of timeliness and suggest that the evaluator should "find out which major audiences intend to make decisions for which they want information and at what times" (Joint committee 1981, p.44). Cousins and Leithwood (1986) emphasize that the timely provision of the evaluation is positively related to utilization.

Substance – Alkin et al (1979) emphasize that the information provided should be "well enough tailored to the users' specific concerns and questions about the program" (p.253). Cousins and Leithwood reported that ...

> "evaluations that reflected knowledge of the context in which the evaluations were to be used, appealed to the preferences of the decision-makers, sought consensus about the evaluation problem, or demonstrated insight into program operations and decision making, were associated with higher levels or use".

Positive findings – Cousins and Leithwood (1986) reported that when evaluation findings were congruent with decision-maker expectations, acceptance and utilization was enhanced. Incongruent findings tended to be ignored. Nevertheless, there are empirical studies which have found that the nature of

findings is unimportant in predicting use (Patton 1977).

Format – several aspects of formats are considered in studies examining information utilization. Stake (1977) distinguished between oral and written communication of results. Other aspects of format are qualitative or quantitative summary of data; the length of the report; single report versus a series of reports prepared at various phases of the program's life cycle. Alkin et al (1979) claim that the format of conveying evaluation information in itself is not a significant determinant or utilization. They claim:

> "What the evaluator has to offer in the way of data and insights may be exactly what the program personnel want or need to know, but may well be ignored or considered useless. Personal preferences must be considered: some people are simply uncomfortable with quantitative data, for example".
>
> (p.254)

Credibility – Credibility is frequently used to describe characteristics of the evaluator. The Joint Committee (1981) uses the term "evaluator credibility" and specifies that the persons conducting the evaluation should be both trustworthy and competent to perform the study. Other evaluators refer through this term directly to the content of the communication. Stufflebeam (1969) interprets this term through the question: "Is the information trusted by the decision maker and those he or she must serve?" Alkin et al (1979), too, indicate that credibility attributed to an evaluator may be limited to certain specific users and to certain limited areas of specialization of the evaluator. It may well be that a certain evaluator has a high level of credibility in some areas and a low level of credibility in other areas. Therefore they, too, refer to credibility of some specific information rather than of persons. Cousins and Leithwood (1986) summarized studies which examined the impact of credibility on evaluation utilization and stated: "when evaluations were perceived by decision-makers as having high face validity, or when they were emphasized as important acitvities, use and the potential for use appeared to be greater". (p.353).

The Process of Conducting Evaluation

During the whole process of conducting a study, the evaluation team should keep in mind issues which may facilitate the utilization of the results. Patton (1985) emphasizes that the evaluation team should ask itself how decisions will affect utilization, whenever questions related to the focus of the study, design, methods, measurement, analysis and reporting emerge.

Planning for utilization – Concern for utilization should be ongoing and continuous from the very beginning of the study.

> Utilization is not something one becomes interested in at the end of an evaluation. By the end of the evaluation the potential for evaluation has been largely determined. From the moment decision-makers and evaluators begin conceptualizing the evaluations, decisions are being made which will affect utilization in major ways.
>
> Patton 1985, p.1748

Orientation toward the User

Evaluation should consider the interests and the needs of the clients. Evaluators should identify the potential users of the study and ascertain what information is likely to be used by them.

Involvement of users – Decision-makers should be involved in making decisions about the evaluation study. People who have stakes in the outcomes of evaluation (the "stakeholders) can play an active role in planning the evaluation study and in determining its major parameters. Being partners of the evaluation study increases the commitment to the utilization of the results.

Multiple clienetle – Various groups may be interested in the evaluation of a particular entity, and the focus of the interest may vary from group to group. For example, the evaluation of a new curriculum may be of interest for teachers, school seupervisors, teacher training institutes, community leaders, parents and the students themselves. Evaluators should display sensitivity to the information needs of the various groups of stakeholders. It does not mean that they have to produce answers to all the questions which representatives of various groups may raise. Limitations of time, resources and staff do not permit all questions to be answered, or full attention to be given to all possible issues. There is a need to determine priorities in evaluation, and this should be done in collaboration with the interested parties.

Training for utilization – Evaluators should assume responsibility for training the decision makers in understanding and utilizing evaluation results. Decision makers should be aware of the significance of the results and of their implications both for short-range and long-range planning. Carrying out such training involves financial and staff time costs. These expenses should be takent into consideration when budgeting for the evaluation program is done.

The Characteristics of the User(s)

The personal characteristics of the designated users, their skills and the organizational characteristics of the institutes in which they work, all have influence on the utilization of the evaluation results. Moreover, even the broader setting in which an institute operates may indirectly effect utilization conditions. Alkin et al (1979) refer to them as "extra organizational factors".

Administrative style of the designated user – A broad variety of administrative and leadership skills may affect the level of utilization. Alkin et al (1979) observed two skills which are directly related to utilization: initiative and organzation skill. By initiative, they refer to a willingness to take charge or to make the first step in a certain direction. It does not mean a mere love of innovation, but rather an openness to new ideas that promise significant benefits. It also includes the ability to estimate probable benefits and risks, and weigh them against each other. Organizational skills refer to the administrator's ability to "get things going".

Organizational factors – Various aspects of the organization climate within an institution may effect the degree of evaluation utilization. Support of cohesive staff may promote utilization just as well as their opposition may hinder the use of a particullar evaluation information. Openness to innovation on behalf of the staff, involvement of the staff in implementing changes, and an open communication pattern among staff members are factors which may contribute to utilization.

Extraorganizational factors – Community interests or the intervention of municipal and governmental factors may be influential in promoting or hindering the utilization of evaluation results. Evaluators have to take into consideration the potenial impact of these factors.

Utility Standards

The publication of the Evaluation Standards by the Joint Committee on Standards for Educational Evaluation (1981) has focused the attention of evaluation experts on the importance of utilization. The Committee defined standards for guiding the work of evaluators and for serving as criteria to evaluate the adequacy of any particular evaluation study. They clustered the standards into four groups: utility, feasibility (to ensure that evaluation should be realistic), propriety (ethical and legal considerations) and accuracy (conveying technically adequate information, i.e. reliable, valid, justified, etc.).

The Utility Standards, as suggested by the name, deal with issues of utilization. The committee specifies eight utility standards as presented in Table 4.2.

Table 4.2 Utility standards

1.	Audience identification (who are involved in, or effected by the evaluation)
2.	Evaluator credibility (competency or persons conducting evaluation)
3.	Information scope and selection (addressing pertinent questions and being responsive to the needs of the audiences)
4.	Valuational interpretation (specifying the value base of judgements)
5.	Report clarity (specifying the evaluation target, purpose, conclusions, etc.)
6.	Report dissemination (the findings should be communicated to the clients and other right-to-know audiences)
7.	Report timeliness (the release of the report should be timely)
8.	Evaluation impact (encouraging follow-through by members of the audiences)

The standards specify the attributes of studies which may enhance utilization.

Most of these standards have been mentioned in the previous sections of this article. The importance of publishing a list of standards is that they provide a comprehensive inventory of issues which should be considered in the planning and conducting of a study.

The authors established a constant format for dealing with each standard and for guiding the reader on how to employ a particular standard. The format consists of the following elements: a formal definition of the standard, and overview (explaining the importance and the rationale), guidelines for action, pitfalls, caveats, an illustrative case, and the analysis of the case. Table 4.3 contains excerpts related to one particular standard.

Table 4.3 Report timeliness

Standard	Release of reports should be timely, so that audiences can best use the reported information.
Overview	Reports are timely when they are delivered to each audience at a time when they can best use the reported information. The potential value of reports that arrive after the occasion for their use is likely to be greatly diminshed.
Guidelines	A. In planning the evaluation, find out which major audiences intend to make decisions for which they want information and at what times. B. Plan backward in time, starting with the targeted delivery dates or the reports, and make realistic projections of what must be done to meet the deadlines, allowing sufficient time to meet unanticipated problems and to meet requests for interim reports.
Pitfalls	A. Promising completion dates without planning a detailed timetable and ensuring that resources are adequate to do the evaluation.
Caveat	A. Consider tradeoffs in order to get reports completed on time. For example, use a smaller sample of the population, shorten interview forms, omit peripheral parts of the evaluation, or shorten the main body of the report. But do not sacrifice

technical accuracy in order to meet a reporting deadline.

Illustrative Case A superintendent of a suburban school district hired an evaluator from a nearby university to evaluate a new teacher aide program for the elementary schools. Because the program was expensive, the school board wanted to evaluate the first year's results before renewing it. The evaluator collected several kinds of data during the year: pupil attitudes, teacher attitudes, aide activities, and job satisfaction. However, the evaluator spent so much time processing and analyzing the data that the report was not finished at the previously agreed upon time. When the board held its final summer meeting to vote on contracts with the aides it had to rely on the superintendent's undocumented impressions of program success in making its decision.

(Joint Committe 1981)

Three of these utility standards will be discussed in detail below.

Information scope and selection

As already indicated, an evaluation study usually has a multiple clientele and various groups focus their interest on different aspects of the program. The evaluators themselves have their own preferences. For example, some are concerned with cost effectiveness, others with the acquisition of values, or with participation in the lesson. Despite personal preferences, the evaluator should strive to answer significant questions of all interest groups.

The evaluator should avoid creating the impression that all questions will be answered. It is necessary to update information requirements through periodic contact with the clients.

Report clarity

Clarity, in the Standards, is described as "explicit and unencumbered narrative, illustrations and decriptions,...well-defined technical terms, tabular or graphic representations, and relevant examples" (p.37).
A formal, written report should be supplemented with other communication techniques, like discussion, television presentation, etc. The authors of the standards present an illustrative case of an evaluation report. A school board received a 250–page report full of ill-defined technical jargon and ambiguities. The recommendations were so vague as to be almost truisms, e.g. "teachers should devote more effort to the needs of the individual students". The tables of the report could not be understood without reference to other sections of the report.

There is little chance that such a report will encourage utilization. The evaluators should have prepared an executive summary containing clear recommendations for action and supplemented by appendices. An example of a clear operational definition recommendation is "reduce the number of supervised extracurricular acitivities in grades 4 through 6 from 25 to no more than 15 and use the savings in teacher time to..."

Report dissemination

There are more problems related to dissemination than merely ensuring that the report should reach the authorized audiences. It may happen that the evaluator has to fight for the right to convey information to those who may use it in a way which is not congruent with the interest of some other group from the multiple clientele. In other cases, the failure to clear the information with key persons before they reach certain audience groups may hinder utilization. Prudence and diplomacy are frequently quite important in the promotion of utilization, and occasionally even more important than skills in evaluation methodology.

The Complexity of Evaluation Utilization

The systematic effort of experts to increase utilization takes place in the face of discouraging reports about the underutilization of evaluation results. Researchers try to explain the reasons for underutilization, suggest ways and produce guidelines which may lead towards growth in evaluation utilization. Rich (1981) believes that underutilization is a result of the lack of communication between evaluators and policy-makers. Improved communication may increase the utilization. This approach is dominant in the Standards of the Joint Committe (1981). They argue for better planning, improved communication, greater involvement of the policy-makers, higher level of accuracy and punctuality, and clearly specified contractual relationships between the evaluators and their clientele. They believe that actions of this type will increase utilization.
The authors whose views have been so far presented believe that strict adherence to recommended procedures, roles, and standards constitutes the key for solving the problem of utilization. A different view is presented by Shapiro (1984) who argued that the reason for underutilization is not the gap between an ideal planning pattern and the actual behavior of the evaluators, but rather the non-adequate conceptualization of "evaluation use". Shapiro claims that evaluation experts usually deal with utilization from the perspective of the rational choice model of decision-making, while in reality this model seldom describes the actual decision making about educational programs. For better understanding of utilization problems, one has to get acquainted with four models of decision-making. These models are: rational choice, bureaucratic politic, organizational process, and cognitive processing.

The rational choice model – The rational choice is frequently misunderstood by those who make reference to it. In this context the term rational applies to the means of attaining certain ends and not to the ends themselves. "The technical definition of rational decision-maker refers to the consistency of behavior or decisions with some social definition of what is good, right or useful". (p.635). This model fits a description of individual decision-making. Rationality does not describe actions of organizations, agencies or institutions, since individuals operating within such a framwork may have conflicting interests.

Bureaucratic-politic model – The bureaucratic model describes decision-making as the product of bargaining and negotiating. Decisions represent an act of mutual adjustment by which a conflict is resolved. Decisions of this type are usually incremental, i.e. they are made step after step, and do not introduce major changes at once.

Organizational processes model – In contrast to the previously described model, the organizational processes model does not describe a decision as a resolution of conflict among individuals, but rather as an act due to consensus among individuals "concerning the need for stability, consistency and the continued existence of the organization... Decisions are consistent with organizational roles and rules designed to ensure the continued existence of the organization" (Shapiro 1984, p.637–38).

Cognitive processing model – The cognitive processing model is adequate for decision-making in the face of an overwhelming complexity which is not amenable to rational calculation or bargaining.

> "In citing Chomsky's model of language acquisition as a process in an environment not sufficiently structured to convey the rules of grammar, Steinbrunner (1974, p.40) asserts that the inferential capacity of the mind to construct grammar rules reveals its ability to impose structure on highly ambiguous data".
>
> Shapiro 1984, p.638

Employing Alternative Decision-making Models

Shapiro (1984) argues that the four models described above do not represent four different realities but rather different aspects of a single reality. Moreover, in every decision, one may detect elements from all four models. The rational choice model may be used for describing the consequences of an evaluation study for the decision-maker alone. It may best fit simple situations, where the consequences of the decision are clear. The bureaucratic model considers the consequences of decision making for various individuals in an organization. It may be used for dealing with those aspects of outcomes which are consensually acceptable. A comparison of treatment versus control group may be not relevant in such a situation. Within the framework of such a model, decision makers can effectively deal with incremental changes and use data which facil-

itates implementing minor changes in a program.

The organizational process model may be useful for selecting between alternative decisions, so that none of them jeopardizes the stability of the organization. Accordingly, decision makers have to restrict their questions and ask "what decision can be made rather than what decisions should be made".(p.642). Within the framework of stable organizations, evaluation results are more likely to be used if they produce surveillance information, i.e. they point out dangerous deviations from organizational routines.

In highly complex situations where decision making is patterned according to the cognitive processing model, evaluation information may be distorted by the user and unconsciously adapted to the decision maker's preconceived inclinations. In such cases the evaluator has little control over the way the information will be interpreted and used by decision makers.

The evaluator has to be aware of the existence of these four types of decision-making patterns. Knowing what decision-making model is likely to be used in a particular situation may help the evaluator to seek data which is relevant to the information needs of his/her clients, and thus also to increase the utilization of the evaluation results.

The Importance of Utilization: An Adversary View

Evaluators as well as administrators attribute great significance to the utilization of knowledge and believe that the value of knowledge is fully determined by the degree of its utilization. This view, however is not fully accepted by all professionals. Buchanan (1985) discusses the pitfalls of overemphasizing the importance of knowledge utilization, and presents the view that utilization cannot be considered the major criteria for determining the value of knowledge. Moreover, overemphasis of knowledge utilization "erodes bases for good practice, and encourages people to take knowledge as more certain and relevant to action than it is" (p.154).

> American Indians know a lot about reptiles that they do not use for purposes of cooking, nor for show and display. Pueblans have an elaborate taxonomy of coniferous trees of no discernible use to them, and Pygmis can distinguish the leaf-eating habits of many species of bats...
>
> ...These so-called primitives are not driven by theoretical or material interest. They explore the world around them to be intelligible. Variations of life and form in nature give people a sense of wonder that feeds on the capacity to name and order things. One never knows: such knowledge may come in handy. Meanwhile, it is a source of pleasure to the individual..."
>
> (p.158)

Knowledge is a source of pleasure even if it is not used as a tool for attaining a certain pre-specified goal.

One of the adverse effects of measuring the merits of an evaluation study by the degree of utilization is that the user is compelled to attribute to the knowledge greater certainty than it actually has. Such certainty is in contradiction to contemporary perspectives about the nature of knowledge. Contemporary views in the field of the philosophy of science, stress the tentative character of knowledge. Philosophers argue that "there is little people can be certain of and call for openness to new data, ideas and systematic criticism" (Buchanan 1985, p.163).

Buchanan emphasizes that in teaching, faith may override knowledge. Devoted teachers tend to disregard the "scientific predictions" about the lack of ability of certain students to learn. The teacher's committment towards students has frequently created better outcomes than what "scientifically valid predictions" have suggested.

The tentative nature of knowledge does not imply that one should not strive to collect data. In social action, the mere fact of gathering information and of paying attention to a variety of phenomena symbolizes values and demonstrates the importance of certain behaviors. Buchanan (1986) refers to court trials and draws an analogy between the act of evaluation and the procedures of a trial. Quoting Tribe (1971), she contends that even if one could increase the accuracy of trial outcomes by using mathematical techniques, it would not be advisable to do so. The procedures of trial symbolize crucial values in our so-

ciety, such as the presumption of innocence, the right to counsel, the privilege against self-incrimination, etc. So also does the process of evaluation symbolize values related to education and to social responsibilities. It also symbolizes the importance of the continuous testing of knowledge and the appropriateness of actual policies.

Evaluation results may motivate people to have second thoughts about the legitimacy of their actions, even if they do not result in observable changes.

Concluding Remarks

This article focused on the experts' concern about the lack of information utilization by evaluation clients and it summarized suggestions and guidelines for increasing utilization. It also touched upon concerns expressed by some experts about the over-utilization of evaluation information and measuring the woth of the evaluation studies in terms of utilization. Alkin et al (1979) wrote about "the graveyard of ignored or disregarded evaluations", and on the other hand Buchanan (1985) expressed concern about ignoring the tentativeness of knowledge, and complained that the problematic concept of knowledge utilization defeats the purpose of knowledge, for it overestimates its certainty.

Emphasizing utilization entails the danger of forcing evaluators to take positions about a controversial issue, even if the empirical evidence does not provide sufficient basis for preferring one or another alternative. If one is conducting an evaluation study, it may be difficult not to provide a clear answer in favor of one particular way of acting.

In this context Buchanan quotes George Eliot

> It is always chilling...to say that you have no opinion to give. And if you deliver an opinion at all, it is mere stupidity not to do it with an air of conviction and well-founded knowledge. You make it your own in uttering it, and naturally get fond of it.
>
> (Eliot 1860/1967 p.23)

But one need not coerce evaluators to take positions. Some of them are inclined to do it, and enjoy doing it. It bestows upon them power, and being listened to increases their self-esteem. An evaluator told the story that he posed a multiple choice type question to fellow evaluators asking them: What

do you find to be the most gratifying reward after completing an evaluation study? Four responses were provided:

1. My colleagues will appreciate the quality of the work
2. I will be rewarded a high honorarium
3. My superiors (or clients) will be satisfied with my work
4. My advice will be implemented

Four out of five evaluators marked answer no. 4. It looks likely that being obeyed is more gratifying than being rewarded or gaining a reputation (Lewy 1980).

The problem of being obeyed is not unique for evaluators. Physicians, too, may sometimes have the problem of convincing people to accept a certain recommended medical treatment. But no physician would confuse the methodology for determining the worth of a certain drug with the methodology of convincing people to take the drug. It seems that in the field of evaluation this distinction is not clear enough.

Once, Dan Stufflebeam wrote about a variety of evaluation categories. He mentioned the category of pseudo-evaluation, which he defined as an evaluation initiated by politicians in order to ensure support for their policy. But one could mention another type of pseudo-evaluation too: namely, that in which the evaluator tries to impose his/her "results" on the decision maker.

It seems that when the results of a study are, by consensus, considered valid by professionals, then these results always effect decision making. Indeed, in reviewing the evaluation literature, one can rarely pinpoint any studies considered by professionals to be consensually valid, but which were then disregarded in the process of decision making. The opposite is true. Quite frequently, the validity of those evaluation studies which affected decision making is highly criticized by professional colleagues.

This is to say that the major problem of evaluation is to clarify what is a valid way of determining the worth of a given entity. The dynamics of the relationship betweeen the evaluator and the decision maker are very important, but they should not overshadow the importance of ensuring the validity of the re-

ported results. It would be quite wrong to effect decision making by research results of dubious validity. In reviewing some recent trends in evaluation theory, one suspects that the question of validity is underemphasized.

Oscar Wilde has a story about an oriental prince whose beloved young wife died very young. He decided to devote his whole life to her memory. Around her coffin, he built a glorious monument and became fully obsessed with beautifying it. For years he worked to improve its beauty, till one day he found that there was only one thing which disturbed the harmony of his edifice. And he ordered the coffin removed.

Are we not moving in the direction of removing from the evaluation dictionary those very concepts which justify its utilization for decision making?

References

Alkin, C.M., 1970, Products for Improving Educational Evaluation, **Evaluation Comments** v.2(3), 1–15.

Alkin, C.M., Daillak R. and White, P., 1979, **Using Evaluation: Does Evaluation Make a Difference.** Beverly Hills, Sage.

Buchanan, M.,1985, What is Irrational about Knowledge Utilization. **Currriculum Inquiry**, v. 15, 153–68.

Cousins, J.B. and Leithwood K.A.,1986, Current Empirical Research on Evaluation Utilization, **Review of Educational Research**, v.56, 331–64.

Davis, Ed., 1981, **Teachers as Curriculum Developers**. Sydney: Allen and Unwin,

Eliot, G., 1967, **The Mill on the Floss**. Oxford: Oxford University Press (original published 1860).

Joint Committee on Standards for Educational Evaluation.,1981, **Standards for Evaluation of Educational Programs, Projects and Materials** New York: McGraw Hill.

Lewy, A., 1977, (ed.) **Handbook of Curriculum Evaluation.** New York: Longmans.

Lewy, A., 1980, Professionals and their Professions, **Phi Delta Kappa** v.13(1), p.16.

Lewy, A. et al, 1981, Decision Oriented Evaluation: The Case of Israel. Philadelphia: Intenational Science Services.

Patton, M.Q. et al., 1977, In Search of Impact: An Analysis of the Utilization of Federal Health Evaluation Research, In Weiss, C.H.(ed.) **Using Social Research in Public Policy Making**, New York: Heath.

Patton, M.Q., 1985, Evaluation for Use. In Husen, T. and Postlewhwaite, T.N. (eds.) **International Encyclopedia of Education,** Oxford: Pergmanon Press.

Rich, R.F., 1981, **Social Science Information and Public Policy Making** San francisco: Jossey Bass.

Shapiro, J.Z., 1984, Conceptualizing Evaluation Use: Implications of Alternative Models of Organizational Decision Making. In Ross, F., Conner, D.G., Altman, C. and Jackson, C. (eds.) **Evaluation Studies, Review Annual**, v.9, Beverly Hills: Sage.

Stake, R., 1977, Responsire Evaluation, In Hamilton, D., (ed.), **Beyond the Numbers Game,** L ›n, Macmillan.

Stufflebeam, D., 1969, Evaluation as Enlightment for Decision Making. In Beatty, W.H., (ed.) **Improving Educational Assessment**, Washington D.C. Association for Supervision and Curriculum Development.

Tribe, L., 1976, Trial by Mathematics: Precision and Ritual in Legal Process. **Harvard Law Review** v.84, 1329–1393.

Subject Index

Academic learning time, 63
Achievement profiles, 32
Achievement tests, 11
Adaptive criteria, 62
Adoption, 53,59
Antecededents, 16
Appeal, studies of, 13
Artistic approach in evaluation, 91–92
Autonomy of teacher, 89

"Black box", 46

Case studies, 56,90,92–94
Causality, 83
Change agent, 53
Change, Scale of, 52
Checklists, 22–25
Children's Television Network, 14
Classroom observation, 62
Clientele, multiple, 112
Collaborative planning, 61
CIPP model, 16
Communication theory, 45
Community involvement, 12, 16
Comparing curricula, 30
Comparing alternative versions, 20
Connoisseurship, 27
Content of curriculum, 26
Course improvement, 3
Course materials, quality, 5
Court decision, 16
Credibility, 111
Critical judgement, 55
Curricular objectives, 22
Curriculum analysis, 55
Curriculum criticism, 25,91
Curriculum decisions, 11
Currriculum Development Centers, 29
Curriculum Implementation, 45
(see Implementation)
Curriculum materials, 18
Curriculum packages, 18
Curriculum structure, 33
Curriculum tryout, 29–31; design of, 30

Data analysis, 55
Data base, 84
Data summary, qualitative, 55
Debugging errors, 33
Decisions oriented evaluation, 5
Description of events, 55
Diffusion milieau, 72
Diffusion models, 53
Diffusion of innovation, 45,49
Diffusion strategies, 53
Discovery learning, 33
Discrepancy, 10,13 (see also Gaps)
Distortion filter, 19
Divisibility, 53

Educational practice oriented program, 51,65
Educational reform, 45
Entry behavior, 67
EPEL 19,20,40
Error analysis, 31
Ethnographic observation, 72
Experimenting society, 17
Experimental design, 20,31,98
Experimental manipulation, 86
Evaluation
  checklist, 22–25
  client, 107,109
  criteria, 27, 30
  influence, cumulative, 108
  product, 110–112
  questions, 6
  report, format of 111
  stakeholders, 99,108
  system, openness of, 74–75
  utilization, definition of, 106
    models of, 118–120
    overemphasis, 120 (see also utilization)

Fidelity approach to implementation, 45
Formative evaluation, 3–8
  and summative evaluation, difference, 38
  multidimensionality of, 3,69
  of implementation, 67–69
  of learners, 4
  roles, 28

Gaps between intended & actual program, 46
(see also transitions)
Goals scales, 4
Goals, selection of, 4

Head Start Planned Variation, 62
Heterogeneous class, 51
Hierarchy of learning units, 18,33
verifying of, 34

IEA,55
Implementation
conditions of, 59
definition of, 45,48
degree of, 47,69
determinants of, 59
evaluation, reasons for, 46
independent & dependent variables, 71
indirect studies of, 55
process of, 46
summative-formative dimensions of, 67
psychometric measures, 88

In-house assembled version, 22
Individualized program, 51,65–66
Information scope, 116
Innovation, 45
Institutional records, 71
Instructional behavior, 89
Instructional film, 30
Instructional materials, 30
Instructional profiles
Intervention, 31
Interviews, 72
Intrinsic evaluation, 22,90
Israel Curriculum Center, 12
Iterations, 18

Joint Committee on Standards,
6,110,11,114,116,118
Judgemental evidence, 23
Judicial evaluation, 90

Leadership, 61
Literary criticism, 27
experts, 28
Long-range goals, 51,73

Mapping sentence, 48
Mastery learning units, 13,18
Mental operation, higher order, 86
Methodologies, mixing, 99
soft, 18
Modular units, 34
Multiple studies, 57

Narrative description, 84
National Science Foundation, 56,77
Naturalistic inquiry, 55
Naturalistic v. rationalistic inquiry,
differences, 82–83
Needs assessment, 10–13, 16
information sources for 11
validity of, 13
New Zealand, 12

Observing single pupil, 95
Opportunity to learn, 89
Outcomes, 65
Outcomes, unintended, 62
Overall rating, 62

Paradigms, qualitative v. quantitative, 81–84
Participation, 63
Personal interaction pattern, 54
Photographic evaluation, 91
Physical properties of curriculum, 26
Positive findings, 110
Precursory behavior of teacher, 67
Predevelopment phase, 10–17
Preordinate design, 28
Process criteria, 62
Program centered studies, 56
Program development, phases of, 8–10
Program utilization, levels of, 59
Prototype evaluation, 17–18
Publication of curriculum, 26
Pupil-materials interaction, 63

Qualitative studies, 90–97
Quantitative studies, 86–90
Questions, higher order, 87

Report clarity, 117

Report dissemination, 117

Sample, non representative, 18
Satisfaction of user, 61
School climate, 16
Self evaluation, 4
Self reporting, 63
Semiprogramed materials, 87
Sequence, 34
Small group teaching, 51
Social system oriented program, 50
Staff development, 61
Stimulated recall, 63
Stimulus-response exchange,21
Story board, 30
Subject matter expert, 13
Subject matter oriented program, 51
Summative evaluation, 3–8
Supplementary exercises, 11
System theory, 45

Talking picture, educational, 22
Targets of evaluation, 58–59
Teacher's behavior, 21
 role, 16
Teacher-pupil relation, 16
Teaching aids, 51
Textbooks, 66
Time sequence, 66
Timeliness, 7,110,115
Topic preference, 15
Transitions, 66 (see also Gaps)
Triangulation, 57
Tryout, 17,29–31
Two steps flow model, 54

Users, 107
 admissible v. non admissible, 107 (see also clients, stakeholoders)
 involvement, 112
 planning for, 112
 training for, 113
Utility standards, 114
Utilization, factors affecting, 109–110

Viewing preference, 14

Author Index

Adar,L., 49,64,66,75
Alkin,C.M., 5,38,40,105–111,113,122,125
Amiad,R., 77
Andersen,D.R., 42
Anderson,R.B., 57,75
Anderson,S., 86,101
Atkin, M., 57,75

Baker,F.B., 100,101
Ball,S., 86,101
Beatty,W.H., 126
Beecham,R., 46,79
Berliner,D.C., 101
Bloom,B.S., 23,40,64,75
Boruch,R., 20,42,98,101
Broughton,A., 72,77
Bruner,J., 33,40
Bryant,J., 42
Buchanan,M., 120,121,122,125

Campebell,D.T., 17,40,99,101
Charters,W.W., 46,75
Comber,L.C., 56,75
Cooley,W.W., 88,101
Cousins,J.B., 106,109,110,111,125
Cronbach,L., 3,4,40
Cunningham,D.J., 8,42

Daillak,R., 125
Davis,E.D., 81,101,105,125
Denham,C., 64,75
Denzin,N.K., 99,101
Devereaux,F.L., 22,40
Dewy,J., 27,40
Dorr-Bremme,D.W., 82,83,86,101
Duby,A., 4,40

Easley, 92,103
Eden,S., 12,40
Eisner,E., 25,27,40,91,101
Eliot,G., 122,125
Eraut,M., 5,22,41,90,101

Fiske,D.W., 99,101
Fox,S., 49,64,66,75
Fraser,Barry, 61,75
Fitz-Gibbon,C., 71,77
Fullen,M., 45,46,59,62,68,75

Gagne,R.M., 33,34,41
Gall,M.D., 25,41,86,101
Glass, G.V. ,76
Goldman,S.V, 102
Goodlad,J., 46,65,75
Guba,E., 55,76,82,83,102
Guttman,L., 48,76

Haberman,A.M.,76
Hall, Gene E., 59,76
Hamilton,D., 72,78
Hampson,D.H., 50,76
Harrison,B. 28,41
Havelock,R.G., 52,59,74,76
Henderson,E.S., 5,42
Hirsch,E.D., 27,41
House, Ernest,R., 57,76
Husen,T., 41,55,76,102

Jenkins,D., 96,102
Jick,Todd D., 98,99,102
Jones,J., 46,75

Katz, Elihu, 54,76
Keeves,J.P., 56,75
Klein,F., 46,65,75
Kliebard,H., 45,76
Kulka,R.A., 102

Leinhardt,G., 70,76,88,102
Leithwood,K.A. 68,76,106,109,110,111,125
Levin,T.,63,77
Lewy,A.,4,6,29,41,55,56,59,77,81,105,108,
123,125
Lieberman,A., 64,75
Liebman,Z., 77
Light,J.A., 33,41
Lincoln,Y.S., 82,83,102
Linn,R., 74,77
Lohnes,P.R., 88,101
Loucks,S.F., 59

Lukas,C., 62,77
Luckey,J.W., 72,77

Martin,J., 102
McDermott,R.P., 95,102
McGaw,B., 57,77
McGrath,J.E., 100,102
McLaughlin,M., 45,61,77
Mann,J.S., 25,41
Markele,S.M., 19,42
Marsh,D.D., 45,61,77
Mayeske,G.W., 90,102
Meyer,M., 15,42
Mielke,K.W., 14,15,42
Montgomery,D.J., 68.76
Morris,L.L., 71,77

Nathenson,M.B., 5,42
Nevo,D., 43,75

Owen,J.M., 59,78

Parlett,M., 72,78
Patton,M.Q., 105,109,111,112,125
Pennington,F.C., 10,11,42
Pomfret,A., 45,46,59,62,75
Postlethwaite,T.N., 31,32,41,42,102

Reynolds,L.G., 33,41
Rezmovic,E.L., 47,69,70,78
Rich,R.F., 118,126
Rieken,H.W., 20,42
Rist,R., 98,102
Rogers,E.M., 53,78
Rosenblum,Y., 21,42
Rosenshine,B., 21,42

Sabar,N., 56,58,78
Sanders,J.R. 8,42
Schafferzik,T., 30
Scheirer,M.A., 46,69,70,78
Schwartz,R., 79
Scriven,M., 3,4,22,42,62,67,78
Seashore-Louis,Karen, 82,84,102
Shapiro,J.Z., 118,119,126
Shoemaker,F., 53,78
Shye,S., 77

Silberstein,M., 37,66,78
Smith, Louis M., 55,78
Sorenson,J.E., 72
Stake,R., 16,38,42,55,78,92,97,103,111,126
Stenhouse,L., 90,103
Stiles, Lindley 46,70
Stuffelbeam,D.L., 15,16,42,110,111,126
Suarez,T., 10,42

Tamir,P., 5,42,75
Templin,P., 91,103
Tribe,L., 121,126
Tyler,R., 25,43

Varenne,H., 102

Walberg,H., 24
Walker,D.F., 30,42
Wang,M., 46,79
Ward,B.A., 101
Ward, William H., 59,79
Weiss,C.H., 86,102,125
Welch, Wayne W., 59,61,79
White,P., 125
White,R.T., 34,42
Wohlleb,C., 62,77
Willis,G., 25,91,102
Wolf,R.L., 90,103

Yoloye,A., 71,79

Zaltman,G., 79
Zuckerman,J.V., 30
Zetterberg,H., 97,103